We propose a general-purpose variational algorithm that forms a natural analogue of Stein variational gradient descent (SVGD) in function space. While SVGD successively updates a set of particles to match a target density, the method introduced here of Stein functional variational gradient descent (SFVGD) updates a set of particle functions to match a target stochastic process (SP). The update step is found by minimizing the functional derivative of the Kullback-Leibler divergence between SPs. SFVGD can either be used to train Bayesian neural networks (BNNs) or for ensemble gradient boosting. We show the efficacy of training BNNs with SFVGD on various real-world datasets.

1 INTRODUCTION

Bayesian inference can be treated as a powerful framework for data modeling and reasoning under uncertainty. However, this assumes that we can encode our prior knowledge in a meaningful manner. Typically, this is done by specifying the prior distribution of the model parameters. However, in machine learning (ML), models potentially consist of millions of parameters with potentially highly complex interactions (e.g., very large neural networks (NNs)). Furthermore, the parameter structure of the models itself is allowed to change during training, e.g., the number of parameter grows when using gradient boosting (GB). This makes defining meaningful prior assumptions for parameter spaces difficult or nearly (practically) infeasible. As we usually do not care about single parameters but the complete resulting function, it seems intuitive to directly express our prior knowledge in hypothesis function space by, e.g., specifying the characteristic length scale, periodicity, or smoothness in general. Fortunately, Bayesian inference can also be formulated in function space. In this case, the prior and posterior distributions are stochastic processes (SPs). The most prominent representative is the Gaussian process (GP), for which the posterior GP can be analytically computed. However, training GPs scale cubically in the number of observations, and the implicit Gaussian likelihood assumption is often violated in reality. In this paper, we introduce Stein functional variational gradient descent (SFVGD). This method provides a general gradient descent method in function space that enables practitioners to train ML models to approximate the posterior SP, assuming certain regularity conditions of the prior SP and the likelihood function hold.

1.1 RELATED WORK

Kernelized Stein Methods These methods combine Stein’s identity with a reproducing kernel Hilbert space (RKHS) assumption. Based on a finite particle set, they can either be used to find the optimal transport direction to match a target density or to estimate the score gradient of the empirical distribution of the particles. The former is called Stein variational gradient descent (SVGD) [Liu & Wang (2016)], and approaches of the latter category are called (non-parametric) score estimators [Zhou et al. (2020)]. Our method internally uses SVGD and forms a natural analogue in function space. Several extensions to SVGD exist, e.g., approaches incorporating second-order information such as [Leviyev et al. (2022)] and the more general matrix-kernel valued approach by [Wang et al. (2019a)]. While these extensions usually outperform SVGD, their computational costs are also higher.

Bayesian Neural Networks (BNNs) Typically, BNNs are NNs with weight priors that are trained via variational inference. The prominent representatives are BNNs using Bayes by Backprop [Blundell et al. (2015)].
We propose a novel natural extension of SVGD in function space (Section 3), which enables the
Another GB-based method is NGBoost proposed by Duan et al. (2019), which directly learns the
(Section 3.3). Using real-world benchmarks, we show that the resulting generator training algorithm is
Training BNNs with our descent is closely related to their method, but while they use a score-based
functional approach, their method is based on approximating the posterior of the model parameters. In contrast to our
approach by Wang et al. (2019b). However, these approaches mainly focus on weight priors, and
empirical findings also only relate to the weight space. In contrast to our work, functional priors can
only be applied if a posterior SP with analytical marginal density exists.

**Functional BNNs (FBNNs)** Sun et al. (2019) proposed to use functional priors to train BNNs. Training BNNs with our descent is closely related to their method, but while they use a score-based
approach for the estimation of the derivative of the Kullback-Leibler divergence $D_{\text{KL}}$ between the
prior SP and the variational SP, we estimate this derivative directly via SVGD. [Wang et al. (2019b)]
also use SVGD for FBNNs but apply SVGD directly to the $D_{\text{KL}}$ between the posterior SP and the
variational SP. Furthermore, their work does not show that this in fact maximizes a lower bound for
the log marginal likelihood. Recently, [Ma & Hernández-Lobato (2021) and Rudner et al. (2021)]
proposed different FBNN approaches that also build upon the results of Sun et al. (2019), but while
their methods are specific to training NN function generators, our method can be used to update a set
of particle functions in general.

**Repulsive Deep Ensembles** Repulsive Deep Ensembles are deep ensembles that incorporate repul-
sive terms in their gradient update, forcing their members’ weights apart. A variety of repulsive
terms are presented in (D’Angelo & Fortuin, 2021) and (D’Angelo et al., 2021), outperforming the
approach by [Wang et al. (2019b)]. However, these approaches mainly focus on weight priors, and
empirical findings also only relate to the weight space. In contrast to our work, functional priors can
only be applied if a posterior SP with analytical marginal density exists.

**GB with Uncertainty** The closest neighbor of our approach applied to GB is the ensemble GB
scheme proposed by [Malinin et al. (2021)], which is based on Bayesian ensembles. In contrast to our
functional approach, their method is based on approximating the posterior of the model parameters. Another GB-based method is NGBoost proposed by [Duan et al. (2019)], which directly learns the
predictive uncertainty; however, prior knowledge can not be taken into account.

1.2 Our Contribution

We propose a novel natural extension of SVGD in function space (Section 3), which enables the
practitioner to match a target SP. This approach can be implemented in a BNN or as GB routine
(Section 3.3). Using real-world benchmarks, we show that the resulting generator training algorithm is
competitive while having less computational costs than the approach of Sun et al. (2019). In contrast
to other existing uncertainty-aware GB algorithms, a GB ensemble, when trained via SFVGD, can
naturally incorporate prior functional information. These versatile applications of our framework are
made possible by providing a unifying view of NNs and GB from a functional analysis perspective.

2 Background

2.1 Supervised ML from a Functional Analysis Perspective

Given a labeled dataset $D \in (\mathcal{X} \times \mathcal{Y})^n \sim P_{xy}$ of $n \in \mathbb{N}$ independent and identically distributed
(i.i.d.) observations from an unknown data generating process $P_{xy}$, a supervised ML algorithm
tries to construct a risk optimal model $f$ under a pre-specified loss $L$. In this case, the function $f$ defines a mapping from the feature space $\mathcal{X}$ to the target space $\mathcal{Y}$. The learning algorithm $\mathcal{L}$ to
construct $f$ is a function mapping from the set of all functions $\bigcup_{n \in \mathbb{N}} (\mathcal{X} \times \mathcal{Y})^n$ to a hypothesis
space $\mathcal{H}$, which is a subset of the set of all functions mapping from $\mathcal{X}$ to the model output space $\mathcal{Y} \subset \mathbb{R}^g$ with $g \in \mathbb{N}$. In order to specify the goodness-of-fit of a function $f$, one can define a loss
function $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}, (y, f(x)) \mapsto L(y, f(x))$, which measures how well the output of a fixed
model $f \in \mathcal{H}$ fits an observation $(x, y) \sim P_{xy}$. In the following, we present supervised ML from a
functional analysis perspective. Here, we fix the observation and associate the loss $L$ with the loss
functional $L_{(x,y)}(f) : \mathcal{H} \to \mathbb{R}, f \mapsto L(y, f(x))$. Based on this loss functional, we can define the risk
functional of a model $f$,

$$R[f] = E_{(x,y) \sim P_{xy}} L_{(x,y)}(f),$$

---

1. If $\mathcal{Y}$ is numeric, $\mathcal{Y} = \mathcal{Y}$. Otherwise, $\mathcal{Y}$ is a numerical encoding of $\mathcal{Y}$. 

---
which measures the expected loss of \( f \), and is used to theoretically identify optimal models. In the following, we will assume that the expectation in Eq. (1) exists and is finite. If we knew the usually unknown data generating process and hence the risk functional, we could update any model \( f \in \mathcal{H} \) in the direction of the steepest descent in \( \mathcal{H} \) w.r.t. \( \mathcal{R} \) by following the negative functional gradient of \( \mathcal{R} \). The negative functional gradient of \( \mathcal{R} \), \( -\nabla_{f} \mathcal{R}[f] \), is itself a mapping from \( \mathcal{X} \) to \( \mathcal{Y} \). For every input location \( x \), this gradient returns the direction in model output space \( \mathcal{Y} \), which points to the locally steepest descent w.r.t. \( \mathcal{R} \). In the following, unless otherwise stated, the functional derivative is taken in the \( L^{2} \) space.

**Proposition 2.1** Assuming sufficient regularity and that \( L(y,f(x)) \) is partially continuously differentiable w.r.t. \( f(x) \), we observe for numeric inputs and model output that

\[
-\nabla_{f} \mathcal{R}[f](x) = -p_{x}(x) \cdot \mathbb{E}_{y \sim p_{y|x}} \frac{\partial L(y,f(x))}{\partial f(x)},
\]

where \( p_{x} \) is the marginal density of \( x \).

The proof is given in A.1.1. In practice, we usually do not know \( p_{x} \), and since our dataset \( D \) is finite, we only have access to \( n \) realizations of \( \frac{\partial L(y,f(x))}{\partial f(x)} \). If the feature space is at least partially continuous, its size \(|\mathcal{X}| = \infty\), and we thus cannot estimate \( -\nabla_{f} \mathcal{R}[f](x) \) without additional assumptions. However, we have access to the functional empirical risk \( \mathcal{R}_{emp,D}[f] := \sum_{(x_{i},y_{i}) \in D} L(x_{i},y_{i})[f] \), for which we assume that it converges in mean to \( \mathcal{R} \) as \( n \to \infty \). Its negative functional gradient can be expressed via the chain rule such that

\[
-\nabla_{f} \mathcal{R}_{emp,D}[f](x) = -\sum_{(x_{i},y_{i}) \in D} \frac{\partial L(y_{i},f(x_{i}))}{\partial f(x_{i})} \cdot \nabla_{f} [f(x_{i})](x),
\]

where \( \nabla_{f} [f(x_{i})] \) is the functional gradient of the evaluation functional of \( f \) at \( x_{i} \), which evaluates to the Dirac delta function \( \delta_{x_{i}} \). However, since we take the functional gradient in \( \mathcal{H} \), \( \nabla_{f} [f(x_{i})] \) becomes the projection of \( \delta_{x_{i}} \) into \( \mathcal{H} \). For example, if \( \mathcal{H} \) is an RKHS with associated kernel \( k \), then \( \nabla_{f} [f(x_{i})](x) = k(x_{i},x) \), i.e., our choice of \( \mathcal{H} \) directly influences the “bumpiness” of \( \nabla_{f} \mathcal{R}_{emp,D}[f] \). Furthermore, we can interpret \( \nabla_{f} \mathcal{R}_{emp,D}[f] \) as a (jump-)continuous functional representation of the dataset \( \partial \mathcal{D}_{L,f} := \{(x_{i},-\frac{\partial L(y_{i},f(x_{i}))}{\partial f(x_{i})})|(x_{i},y_{i}) \in D\} \subset ((\mathcal{X} \times \mathcal{Y})^{n}) \), which also implicitly defines a learner. In the following, we show how two core supervised ML algorithms (gradient boosting and neural networks) naturally incorporate this functional gradient while training.

**Gradient Boosting (GB)** For GB [Friedman, 2001], the situation is usually reversed, and we choose a (base) learner \( \mathcal{I}_{b} \) that implicitly defines \( \mathcal{H} \) and with which we fit a model to the data set \( \partial \mathcal{D}_{L,f} \). GB uses these approximations of the negative functional gradient of the empirical risk to successively update a model \( f^{[t]} \) such that

\[
f^{[t+1]} = f^{[t]} + \eta^{[t]}b^{[t]} \quad \text{with} \quad b^{[t]} = \mathcal{I}_{b}(\partial \mathcal{D}_{L,f^{[t]}}),
\]

where \( \eta^{[t]} \in \mathbb{R}_{>0} \) is the learning rate and possibly depends on the iteration \( t \in \mathbb{N} \). For further details see Appendix A.2.

**Neural Networks (NNs)** If \( f \) is an NN with parameters \( \phi \), then the parameter gradients w.r.t. the empirical risk functional needed for backpropagation can be obtained via the chain rule such that

\[
\nabla_{\phi} \mathcal{R}_{emp,D}[f] = \int_{\mathcal{X}} \nabla_{f} \mathcal{R}_{emp,D}[f](x) \cdot \nabla_{\phi} f(x) dx = \sum_{(x_{i},y_{i}) \in D} \frac{\partial L(y_{i},f(x_{i}))}{\partial f(x_{i})} \cdot \nabla_{\phi} f(x_{i}),
\]

where the second equality holds, since here we do not restrict \( \mathcal{H} \), i.e., \( \nabla_{f} [f(x_{i})] = \delta_{x_{i}} \).

However, these procedures only assure that we can find an optimal model \( f \in \mathcal{H} \) w.r.t. \( \mathcal{R}_{emp} \), which does not imply that \( f \) is optimal w.r.t. \( \mathcal{R} \). In practice, we tune the hyperparameters of the algorithms – i.e., use data withheld from learning for subsequent model selection – and apply early stopping to find a model \( f \) approximately optimal w.r.t. \( \mathcal{R} \).
2.2 Stochastic processes

In this section, we will shortly introduce stochastic processes (SPs) that can be used to represent distributions over functions and thereby allow us to express the uncertainty of models independent of their specific parameter structure. We will regard $\mathcal{X}$ as an index set and let $\mathcal{Y}$ be a measurable space with $\sigma$-algebra $\mathcal{G}$ on the state space $\mathcal{Y}$. For $x \in \mathcal{X}$ and a given probability space $(\Omega, \mathcal{F}, \mathbb{P})$ based on the sample space $\Omega$, $\mathcal{F}$ is a $\sigma$-algebra on $\Omega$ and probability measure $\mathbb{P}$. Let $Q(x)$ be a random variable projecting from $\Omega$ to $\mathcal{Y}$. An SP $Q$ is the family $\{Q(x): x \in \mathcal{X}\}$ of all random variables $Q(x)$ [Lamperti 1977]. With this, we can define a sample function $f_\omega: \mathcal{X} \to \mathcal{Y}$, $x \mapsto Q(x)(\omega)$ for a fixed $\omega \in \Omega$. Often, it is easier to look at SPs from this sample function view: For every $A \in \mathcal{F}$, a set of functions $\{f_\omega: \omega \in A\}$ with an associated measure $\mathbb{P}(A)$ can be identified – i.e., SPs define a distribution over functions projecting from $\mathcal{X}$ to $\mathcal{Y}$. For a finite index set $\mathcal{X} := x_{1:m} \in \mathcal{X}^m$, we denote the finite-dimensional marginal joint distribution over function values $\{Q(x_1), \ldots, Q(x_m)\}$ as $Q_\mathcal{X}$. In the following, we assume that for every $Q_\mathcal{X}$ exists a corresponding density function $p_{Q_\mathcal{X}}: \mathcal{Y}^m \to \mathbb{R}_{\geq 0}$, $f^\mathcal{X} \mapsto p_{Q_\mathcal{X}}(f^\mathcal{X})$, where $f^\mathcal{X} := (f(x_1), \ldots, f(x_m))$ are the function values at $x_{1:m}$ based on a sample function $f$ where we suppressed the $\omega$ to ease the following notation. We will denote the associated functional to this density function with $p_{Q_\mathcal{X}}[f]$.

The $D_{KL}$ is a measure of distance between two distributions over the same probability space. Since SPs are distributions over functions, the $D_{KL}$ can also be used for distances between two SPs. Unfortunately, computing this quantity is non-trivial [Matthews et al. 2015]. However, for two consistent and ergodic SPs $P$ and $Q$, i.e., $Q$ and $P$ can be characterized by marginals over all finite index sets (e.g., GPs), Sun et al. [2019] showed that the $D_{KL}$ between these SPs can be solely expressed in terms of their marginals, i.e.,

$$D_{KL}(Q||P) = \sup_{m \in \mathbb{N}, \mathcal{X} \in \mathcal{X}^m} D_{KL}(Q_\mathcal{X}||P_\mathcal{X}).$$  

(6)

This expression enables us to find a differentiable distance measure between two stochastic processes.

2.3 Stein variational gradient descent

SVGD [Liu & Wang 2016] is a variational Bayesian method. Variational methods can be used to approximate the generally intractable posterior density of a continuous random variable $\theta$

$$p_{\theta|\mathcal{D}}(\theta) = \frac{p_{\mathcal{D}|\theta}(\mathcal{D}|\theta)p_{\theta}(\theta)}{\int p_{\mathcal{D}|\theta}(\mathcal{D}|\theta)p_{\theta}(\theta)d\theta},$$  

(7)

where $p_{\mathcal{D}|\theta}$ and $p_{\theta}$ are the likelihood and the prior density function, respectively. SVGD tries to match the posterior $p_{\theta|\mathcal{D}}$ with a density $q$ represented via a fixed number $r \in \mathbb{N}$ of pseudo-samples – so-called particles – and iteratively updates them by minimizing $D_{KL}(q||p_{\theta|\mathcal{D}}) = \int q(\theta) \log \frac{q(\theta)}{p_{\theta|\mathcal{D}}(\theta)} d\theta$.

In an RKHS with associated kernel $k$, the optimal update direction is found by considering the negative functional derivative

$$- \nabla f_{D_{KL}(q||p_{\theta|\mathcal{D}})} = \mathbb{E}_{\theta \sim q} \left[ \nabla_{\theta} \log p_{\theta|\mathcal{D}}(\theta) k(\theta, \cdot) + \nabla_{\theta} k(\theta, \cdot) \right],$$

(8)

where $T(\theta) = \theta + f(\theta)$, and $q_{\mathcal{D}}$ is the density of $\theta' = T(\theta)$ when $\theta \sim q$. We can estimate this functional gradient based on the particles in an unbiased manner, as we are able to evaluate the score function of $p_{\theta|\mathcal{D}}$ (i.e., $\nabla_{\theta} \log p_{\theta|\mathcal{D}}$), although $\log p_{\theta|\mathcal{D}}$ might be intractable.

3 Stein functional variational gradient descent

In this section, we develop a functional version of SVGD which we will call Stein functional variational gradient descent (SFVGD). While SVGD can be used to approximate the posterior distribution of a continuous random variable, SFVGD can be applied when we are interested in the posterior SP $P_{\mathcal{D}|\theta}$ defined by its Radon-Nikodym derivative [Schervish 1995] w.r.t. the prior SP $P_{\theta}$,

$$\frac{dP_{\mathcal{D}|\theta}}{dP_{\theta}}[f] = \frac{p_{\mathcal{D}|\theta}[f]P_{\theta}}{\int p_{\mathcal{D}|\theta}[f]dP_{\theta}[f]},$$

(9)

where $p_{\mathcal{D}|\theta}$ is the likelihood functional, which measures how likely it is to observe $\mathcal{D}$, given a sample function $f$. In the following, we assume that the posterior $P_{\mathcal{D}|\theta}$ exists and also that it is an ergodic and consistent SP. Analogously to SVGD, we try to approximate $P_{\mathcal{D}|\theta}$ with a distribution $Q$ represented by pseudo-samples. However, for SFVGD, these particles are now functions.
3.1 OBJECTIVE FUNCTION

Since analytical solutions for the differential Eq. (9) only exist in special cases (e.g., if the prior \( P_f \) is a GP and the likelihood is also Gaussian), we use the \( D_{KL} \) between two SPs to formulate an optimization objective. More specifically, the goal of our framework is to construct an approximating measure \( Q^* \) for which it holds that

\[
Q^* \in \arg\min_{Q \in \mathcal{Q}} D_{KL}(Q \| P_f|\mathcal{D}),
\]

(10)

where \( \mathcal{Q} \) is the set of representable variational posterior processes. Here, we represent \( Q \) via \( r \in \mathbb{N} \) sample functions \( f_1, \ldots, f_r \) from \( Q \), which act as pseudo-samples and which we also call particle functions. It can be shown (Matthews et al., 2015) that minimizing Eq. (10) is equivalent to maximizing the functional evidence lower bound (ELBO) \( \mathcal{L}_D \), i.e.,

\[
Q^* \in \arg\max_{Q \in \mathcal{Q}} \mathbb{E}_{f \sim Q} \left[ \mathbb{E}_{\mathcal{D} \sim Q} [\ell(\mathcal{D} | f)] - D_{KL}(Q \| P_f) \right],
\]

(11)

where \( \ell(\mathcal{D} | f) := \log p(\mathcal{D} | f) \). The advantage of formulation (11) over (10) is that Eq. (11) only depends on known quantities. In the following, we apply Eq. (6), i.e., the results of Sun et al. (2019) regarding the \( D_{KL} \) of ergodic and consistent SPs, yielding

\[
Q^* \in \arg\max_{Q \in \mathcal{Q}} \inf_{m \in \mathbb{N}, \mathcal{X} \in \mathcal{X}^m} \mathbb{E}_{f \sim Q} [\mathbb{E}_{\mathcal{D} \sim Q} [\ell(\mathcal{D} | f)] - D_{KL}(Q_{\mathcal{X}} \| P_{f|\mathcal{X}})],
\]

(12)

where \( \mathcal{D}_s \) is a random subsample of size \( |\mathcal{D}_s| = s \) drawn from \( \mathcal{D} \), \( \mathcal{X}_{\mathcal{D}_s} \) are the associated feature vectors of \( \mathcal{D}_s \), and \( \mathcal{X}_M = [\mathcal{X}_1, \ldots, \mathcal{X}_M]^\top \in \mathcal{X}^M \) are \( M \) stacked random feature vectors drawn from a sampling distribution \( C_{\mathcal{X}} \) with support \( \mathcal{X} \). If \( \mathcal{X} \) is bounded, Sun et al. (2019) proposes a uniform distribution for \( C_{\mathcal{X}} \). It has been shown in Sun et al. (2019) for \( \mathcal{D}_s = \mathcal{D} \) and \( M > 1 \) that \( \mathcal{L}_{\mathcal{D}_s}[\mathcal{X}_{\mathcal{D}_s}, \mathcal{X}_M](Q) \) is a lower bound for the log marginal likelihood \( \log p(D) \), i.e., the maximization in Eq. (13) implies the minimization in Eq. (10). Although, as noted by Burt et al. (2020), if \( Q \) is a parametric family, the objective is ill-defined, we did not encounter any problems in practice. Also, we could straightforwardly use the grid functional \( D_{KL} \) proposed by Ma & Hernández-Lobato (2021), which fixes some of these theoretical shortcomings. However, note that SFVGD itself does not assume \( Q \) to be parametric.

3.2 FUNCTIONAL DERIVATIVE OF THE OBJECTIVE

When using conventional gradient descent methods, we want to apply a map to update the parameters of our model such that our loss is reduced. In SFVGD, we proceed in a similar manner but update functions towards a loss-minimizing direction. A map that takes a function as an argument and returns another function is called an operator. Hence, we want to express how our objective value Eq. (13) changes when an operator \( F : \mathcal{H} \rightarrow \mathcal{H}, f \rightarrow f_F \) is applied to every \( f \sim Q \). This means that the objective value changes with \( F \) such that

\[
\mathcal{L}_{\mathcal{D}_s, \mathcal{X}}(Q|\mathcal{T}) = \mathbb{E}_{f \sim Q|\mathcal{T}} \left[ \ell(\mathcal{D}_s | f) \right] - D_{KL}(Q|\mathcal{T}_H \| P_{f|\mathcal{T}_H}),
\]

(14)

where \( T(f) = f + F(f) \) and \( Q|\mathcal{T} \) is the distribution of \( f = T(f) \) when \( f \sim Q \). Naturally, we are interested in the functional derivative of Eq. (14) w.r.t. to \( F \), since this gives us the direction of the steepest ascent in operator space regarding the functional ELBO. However, in order to make our computations tractable, we must limit the space of feasible operators:
We call \( \nabla \) where we assumed that there exists a log-likelihood function \( \ell \). Thus, \( \nabla \) based on the particle functions \( f \). For an evaluation-only dependent operator \( Q \), \( \nabla \) of SFVGD is that by updating every particle function \( f \). We call such an operator “evaluation-only dependent”.

Thus, \( F \) does not depend on derivatives of \( f \) (which is not a restriction, since we only assumed \( f \) to be continuous); we can also treat \( F \) as a construction rule of \( F_X \) for arbitrary \( m \) and \( X \). Now, we can state the functional gradient of the objective functional w.r.t. an evaluation-only dependent operator \( F \), for \( X = [X_D, X_M] \) and \( X = X_D \).

\[
\nabla_F L_{D, X}(Q[T], f) = \nabla_F \mathbb{E}_{f \sim Q[T]} \ell[D, f] - \nabla_F D_{KL}(Q[T] \| F_X)
\]

(15)

where we assumed that there exists a log-likelihood function \( \ell : \mathcal{X} \times \mathcal{Y}^s \times \mathcal{Y}^s \rightarrow \mathbb{R} \) such that \( \ell[D, f] = \ell(D, f^X) \) for every \( D \). If we set \( F = 0 \), then \( T \) becomes the identity operator, i.e. \( Q[T] = Q \). Since we want to iteratively update our particle functions, we must only consider small perturbations around \( F = 0 \).

**Proposition 3.1** For an evaluation-only dependent operator \( F \), the functional derivative of the functional ELBO at \( F = 0 \) evaluated for a function \( f \)

\[
\nabla_F L_{D, X}(Q[T], f) = \mathbb{E}_{f \sim Q_X} \nabla_Y \ell(D, f^X) \cdot \delta_X(f^X) \cdot \left[ \delta_X(\cdot), \ldots, \delta_X(\cdot) \right]^T
\]

(16)

where we assume that \( \mathcal{H}_Y \subset \{ f : \mathcal{Y}^s+M \rightarrow \mathcal{Y}^s+M \} \) is an RKHS with associated kernels \( k_Y \).

The proof is given in [A.1.2] where we also show the following corollary.

**Corollary 3.1.1** For an evaluation-only dependent operator \( F \), the functional derivative of the functional ELBO at \( F = 0 \) evaluated for a function \( f \)

\[
\nabla_F L_{D, X}(Q[T], f) = \mathbb{E}_{f \sim Q_X} \nabla_Y \ell(D, f^X) \cdot k_Y(f^X) \cdot \left[ \delta_X(\cdot), \ldots, \delta_X(\cdot) \right]^T
\]

(17)

where we assume that \( \mathcal{H}_Y \subset \{ f : \mathcal{Y}^s+M \rightarrow \mathcal{Y}^s+M \}, \mathcal{H}_X \subset \{ f : \mathcal{Y}^s \rightarrow \mathcal{Y}^s \} \) are RKHSs with associated kernels \( k_X, k_Y \), respectively.

We call \( \nabla_F L_{D, X}(Q[T], f) \) the Stein functional variational gradient operator. It inherits its name from SVGD, which internally is used to find the functional derivative of the \( D_{KL} \) term. The key idea of SFVGD is that by updating every particle function \( f \sim Q \) via functional gradient descent in the direction of \( \nabla_F L_{D, X}(Q[T], f) \), we carry out a gradient step in the distribution space. This increases the current overall functional ELBO value we want to maximize by pulling \( Q \) closer to \( Q^* \) and consequently also closer to the true posterior stochastic process \( P_f[D] \).

### 3.3 Algorithms

Based on the particle functions \( f_1, \ldots, f_r \), we can find an estimator of Eq. \[16\]

\[
\nabla_F L_{D, X}(Q[T], f) = \frac{1}{r} \sum_{i=1}^r \nabla_f \ell(D, f_i^X) \cdot \delta_X(f_i^X) \cdot \left[ \delta_X(\cdot), \ldots, \delta_X(\cdot) \right]^T
\]

(18)
We choose for $\lambda \in \mathbb{R}_{\geq 0}$. Furthermore, if we set $\lambda = 1$, the estimator becomes an unbiased estimator of Eq. (16). Since $\mathcal{L}_{\mathcal{D}, f}$ is a lower bound of the log marginal likelihood $\log p(D)$, it would be preferable to update the particle functions via $\tilde{\nabla}_f \mathcal{L}_{\mathcal{D}, f}(Q_{[T]}(\cdot)) |_{T=0}$. However, the major computation bottleneck in Eq. (18) is the calculation of the score gradient $\nabla_{\delta f_i} p_{f_i}(\delta f_i \delta f_i)$ for all particle functions $f_i, i = 1, \ldots, r$ evaluated at $\mathbf{X}$. For example, if $P_f$ is a GP, then the costs of computing $\nabla_{\delta f_i} p_{f_i}(\delta f_i \delta f_i)$ are $O((s + M)^2)$. In addition, the computation of all kernel values $k_Y(f_i, f_j \delta f_i)$, $i = 1, \ldots, r, j = 1, \ldots, r$ required in Eq. (18) costs $O(r^2)$. However, this is usually small compared to the cost of computing the score gradient for the functional prior.

We choose for $M$ a small constant number, since $\mathcal{L}_{\mathcal{D}, [\mathbf{X}_p, \mathbf{X}_n]}$ is a lower bound for the log marginal likelihood $\log p(D)$ for $M > 1$, and we set $r$ to a number of particle functions that can represent the posterior SP reasonably well. Thus, we are interested in estimating $\tilde{\nabla}_f \mathcal{L}_{\mathcal{D}, f}(Q_{[T]}(\cdot)) |_{T=0}$ with mini-batches. In principle, an unbiased estimate of $\ell(D, f_1 \delta f_i)$ is $n / s \cdot (\ell(D_s, f_i \delta f_i))$, which suggests that $\lambda = s / n$. Although (in general) $\mathcal{L}_{\mathcal{D}, f}$ is not a lower bound of $\log p(D)$, we found in a practice setting that $\lambda$ to $s / n$ still results in reasonable performance. However, our theoretical framework gives the reassuring guarantee that if we use full-batch training, we would, in fact, maximize a lower bound of $\log p(D)$. In the following, we present two algorithms, namely Stein functional variational NNs and Stein functional variational gradient boosting (A3.1), based on the estimated Stein functional variational gradient – i.e., they depend on the score gradient of the functional prior evaluated at $\mathbf{X}$. If there exists no analytical score gradient, we can use a score gradient estimator, as suggested in Sun et al. (2019). This only requires function samples of the prior process evaluated at $\mathbf{X}$, but estimating the score gradient is usually computationally expensive (Zhou et al., 2020). Since our approach builds upon SVGD, there exists an additional approach in our framework based on a gradient-free SVGD (Han & Liu, 2018) that only requires the evaluation of the marginal densities of the prior process.

Stein Functional Variational Neural Network (SFVNN) Sun et al. (2019) proposed to train neural networks (NNs) acting as function generators with the negative functional ELBO as loss, which they call Bayesian Functional Variational Neural Networks (BFVNNs). Such a function generator can be modeled via an NN with stochastic weights, which can be represented as a differentiable function $g : \mathcal{Z} \rightarrow \mathcal{Y}, z \mapsto g(z)$, where $z \in \mathcal{Z}$ consists of the deterministic input $x$ and stochastic...
inputs, i.e., we can model $z$ as a random variable $z \sim p(z|x)$. These NNs are applicable as long as the reparameterization trick (Kingma & Welling, 2014) can be used, i.e., there exists a random variable $\xi \in \Xi$ with $\xi \sim p(\xi)$ and a differentiable function $h_\phi : \mathcal{X} \times \Xi \to \mathcal{Y}$ parametrized by $\phi$ such that $h_\phi(x, \xi) \sim p(z|x)$. With this, we can sample a function by sampling $\xi \sim p(\xi)$ and defining $f_\xi : \mathcal{X} \to \mathcal{Y}$, $\mathbf{x} \mapsto g(h_\phi(\mathbf{x}, \xi))$. In this case, we can write the gradient of Eq. 14 w.r.t. $\phi$ as

$$
\nabla_\phi \mathcal{L}_{D,\mathcal{X}}(Q_{\mathcal{T}}) = \mathbb{E}_{\zeta \sim p(\zeta)} \left[ \nabla_{f_\xi} \mathcal{L}_{D,\mathcal{X}}(Q_{\mathcal{T}}) \right] - \mathbb{E}_{\zeta \sim p(\zeta)} \left[ \left( \nabla_{f_\xi} \log p_{Q_{\mathcal{X}}}(f_\xi) - \nabla_{f_\xi} \log p_{P_{\mathcal{X}}}(f_\xi) \right) \nabla_{f_\xi} \log p_{Q_{\mathcal{X},\phi}}(\mathcal{F}_\xi) \right].
$$

This is also the result obtained in Sun et al. (2019), where they then use a score estimator (namely, the spectral stein gradient estimator (SSGE; Shi et al., 2018)) to approximate $\nabla_y \log p_{Q_{\mathcal{X}}}(f_\xi)$. SSGE estimates the score gradient in an RKHS, i.e., the entropy gradient $\nabla_{f_\xi} \log p_{Q_{\mathcal{X}}}(f_\xi)$ and the cross entropy gradient $\nabla_{f_\xi} \log p_{P_{\mathcal{X}}}(f_\xi)$ are taken in different functional spaces. Our SFVNN is based on the parameter gradient

$$
\nabla_\phi \mathcal{L}_{D,\mathcal{X}}(Q_{\mathcal{T}}) = \mathbb{E}_{\zeta \sim p(\zeta)} \left[ \nabla_{f_\xi} \mathcal{L}_{D,\mathcal{X}}(Q_{\mathcal{T}}) \right]_{F=0} - \mathbb{E}_{\zeta \sim p(\zeta)} \left[ \nabla_{f_\xi} \log p_{Q_{\mathcal{X},\phi}}(\mathcal{F}_\xi) \right]_{F=0},
$$

where we use the general Stein functional variational gradient from Eq. 16. In contrast to Sun et al. (2019), we thereby directly take the functional gradient of the $D_{KL}$ term in an RKHS, and our score gradients of the prior process are also subject to the implicit kernel smoothing.

**Runtime comparison between SVFNN and FVBN** While FVBN scales as $\mathcal{O}(r^3 + r^2(s + M))$ (because of SSGE), our approach scales only quadratically in $r$ (because of SVGD), allowing for a larger number of sample functions (see Appendix B).

### 3.4 Illustrative Example

![Illustrative example of SFVGB](image)

Figure 1: Illustrative example of SFVGB. The points in (a) represent the given data points, and the dashed lines represent all $x$ values that define the marginal prior density in (c,d) w.r.t. the prior GP shown in (b). The arrows in (c) show the resulting SFVGD gradients.

Given three sample functions and two data points $\{(0.75, 1.0), (2.0, 0.0)\}$, we want to approximate the posterior GP (Figure 1h) w.r.t. the prior GP shown in Figure 1b and a Gaussian likelihood via SFVGB. Hence, we also sample a necessary measure point $x_M = 1.5$. The resulting three-dimensional marginal density is defined by the prior GP. SVGD gives us the optimal update direction for the sample function values to fit this marginal density (Figures 1c, 1d). We fit a kernel ridge regression to these directions after adding the log likelihood gradients at the two data points (Figure 1e). The resulting updated function samples after this SFVGD step can be seen in Figure 1f and the converged function samples in Figure 1g. Qualitatively comparing these converged function samples in Figure 1g with sample functions from the exact posterior GP in Figure 1h reassures that we are able to approximate the posterior GP in this toy example reasonably well.
Table 1: Comparison of different methods (columns) on small benchmark data sets (rows) using the average NLL (smaller is better) and RMSE over 10 train-test data splits with standard deviation in brackets. The best performing method for each data set is highlighted in bold.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>SFVNN (0.17)</th>
<th>FVBNN (0.04)</th>
<th>BNN (0.12)</th>
<th>GP (0.14)</th>
<th>SFVNN (0.20)</th>
<th>FVBNN (0.19)</th>
<th>BNN (0.40)</th>
<th>GP (0.25)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Airfoil</td>
<td>2.10</td>
<td>2.29</td>
<td>2.62</td>
<td>2.50</td>
<td>1.82</td>
<td>1.97</td>
<td>2.40</td>
<td>2.77</td>
</tr>
<tr>
<td>Concrete</td>
<td>2.59</td>
<td>3.07</td>
<td>3.25</td>
<td>3.06</td>
<td>4.58</td>
<td>4.64</td>
<td>6.18</td>
<td>5.13</td>
</tr>
<tr>
<td>Diabetes</td>
<td>5.42 (0.08)</td>
<td>5.49 (0.03)</td>
<td>5.41 (0.04)</td>
<td>6.19 (0.38)</td>
<td>54.57 (3.74)</td>
<td>57.1 (2.48)</td>
<td>52.7 (2.88)</td>
<td>57.45 (6.6)</td>
</tr>
<tr>
<td>Energy</td>
<td>0.62 (0.10)</td>
<td>0.70 (0.09)</td>
<td>2.26 (0.32)</td>
<td>2.38 (0.05)</td>
<td>0.44 (0.05)</td>
<td>0.43 (0.08)</td>
<td>2.37 (0.65)</td>
<td>2.34 (0.23)</td>
</tr>
<tr>
<td>ForestF</td>
<td>2.38 (0.44)</td>
<td>1.84 (0.05)</td>
<td>1.83 (0.05)</td>
<td>4.65 (0.45)</td>
<td>1.76 (0.31)</td>
<td>1.51 (0.07)</td>
<td>1.51 (0.08)</td>
<td>1.56 (0.08)</td>
</tr>
<tr>
<td>Wine</td>
<td>1.96 (1.45)</td>
<td>1.47 (1.07)</td>
<td>-0.03 (0.07)</td>
<td>-0.04 (0.06)</td>
<td>0.11 (0.02)</td>
<td>0.14 (0.02)</td>
<td>-0.21 (0.03)</td>
<td>0.16 (0.03)</td>
</tr>
<tr>
<td>Yacht</td>
<td>1.06 (0.30)</td>
<td>1.11 (0.24)</td>
<td>1.35 (0.19)</td>
<td>2.86 (0.15)</td>
<td>0.67 (0.26)</td>
<td>0.61 (0.25)</td>
<td>0.96 (0.28)</td>
<td>3.95 (1.03)</td>
</tr>
</tbody>
</table>

Mean rank: SFVNN = 1.86, FVBNN = 2.43, BNN = 2.57, GP = 3.14

Table 2: Comparison of different methods (columns) on large benchmark data sets (rows) using the average NLL (smaller is better) and RMSE over 10 train-test data splits with standard deviation in brackets. The best performing method for each data set is highlighted in bold.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>SFVNN (0.04)</th>
<th>FVBNN (0.03)</th>
<th>BNN (s.n)</th>
<th>GP (s/n)</th>
<th>SFVNN (1.26)</th>
<th>FVBNN (0.9)</th>
<th>BNN (0.03)</th>
<th>GP (0.01)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPU</td>
<td>4.73</td>
<td>4.80</td>
<td>4.73</td>
<td>27.6</td>
<td>27.2 (1.26)</td>
<td>27.6 (0.9)</td>
<td>25.6 (0.67)</td>
<td></td>
</tr>
<tr>
<td>NavalT</td>
<td>-6.51 (0.06)</td>
<td>-6.85 (0.08)</td>
<td>-5.03 (0.24)</td>
<td>-6.85 (0.00)</td>
<td>1.70E-4 (2.5E-5)</td>
<td>1.94E-4 (3.5E-5)</td>
<td>5.60E-4 (5.6E-5)</td>
<td></td>
</tr>
<tr>
<td>NavalC</td>
<td>-6.53 (0.10)</td>
<td>-6.41 (0.05)</td>
<td>-6.44 (0.11)</td>
<td>-6.44 (0.11)</td>
<td>1.35E-4 (1.1E-4)</td>
<td>2.39E-4 (3.5E-5)</td>
<td>2.15E-4 (3.9E-5)</td>
<td></td>
</tr>
<tr>
<td>Protein</td>
<td>2.85 (0.01)</td>
<td>2.87 (0.01)</td>
<td>2.96 (0.01)</td>
<td>-6.65 (0.08)</td>
<td>4.19 (0.04)</td>
<td>4.27 (0.04)</td>
<td>4.65 (0.05)</td>
<td></td>
</tr>
<tr>
<td>VideoMem</td>
<td>11.39 (0.10)</td>
<td>11.39 (0.10)</td>
<td>11.4 (0.03)</td>
<td>11.4 (0.03)</td>
<td>21.19 (49.10)</td>
<td>21.000 (2320)</td>
<td>21.800 (788)</td>
<td></td>
</tr>
<tr>
<td>VideoTime</td>
<td>2.29 (0.05)</td>
<td>2.53 (0.36)</td>
<td>2.86 (1.02)</td>
<td>2.86 (1.02)</td>
<td>2.51 (0.16)</td>
<td>3.14 (0.88)</td>
<td>3.83 (2.08)</td>
<td></td>
</tr>
</tbody>
</table>

Mean rank: SFVNN = 1.25, FVBNN = 2.17, BNN = 2.58, GP = 2.58

4 Benchmark Study

We further investigate the competitiveness of our approach using its neural network variant (SFVNN) with its closest neighbor, the functional variational Bayesian neural network (FVBNN) from Sun et al. (2019). We also include one well-established BNN baseline (Blundell et al., 2015) and the standard Gaussian Process for the small data sets (where analytical computation is feasible). For results from SVFGB we refer to Section A.3.2 in the Appendix. For most of the datasets, however, the NN provides a better fit to the data. Further details and a contextual bandits experiment can be found in Appendix A.5.

Data and experimental setup All data sets are standardized prior to model fitting and split into 90% training data and 10% test data. For the comparisons, this splitting process is repeated 10 times based on 10 different splits to also evaluate the variability of each method. Further details on data sets, data set-specific pre-processing, and their references can be found in the Appendix.

Details on methods and comparisons In order to provide a fair comparison between methods, we reproduce the best results reported by Sun et al. (2019) for FVBNN and BNN, and also use the same hyperparameters for our method except that while Sun et al. (2019) use $\lambda = 1$, we set $\lambda = s/n$. Details for each procedure are given in the Appendix. We compare methods based on the negative log-likelihood (NLL) and the root mean squared error (RMSE) on each test data set and calculate the mean and standard deviation across all 10 data splits.

Results Results are summarized in Tables 1 and 2 indicating that SFVNN is competitive with other existing approaches for both small and large data sets. As the two functional approaches (SFVNN, FVBNN) optimize the same objective, we would expect them to perform similarly, which is confirmed by the results. We further observe that a weight space approach (the BNN) seems to work better than the functional approaches for a few datasets (in particular, for the Wine dataset, where this is expected as the outcome is of discrete nature).

5 Conclusion

We introduced a novel gradient descent in distribution space that allows us to update a set of particle functions in a general manner to resemble sample functions from a target process. SFVNN was found to be competitive with or to outperform FVBNN while having less computational costs.
REFERENCES


A APPENDIX

A.1 PROOFS

A.1.1 FUNCTIONAL DERIVATIVE OF THE RISK FUNCTIONAL

Assuming that $p(x, y)L_{(x,y)}[f] \in L^1$ using Fubini’s theorem, we find that

$$
\mathcal{R}[f] = E_{(x,y) \sim p_{xy}} L_{(x,y)}[f] = \int_{X \times Y} p(x, y) L_{(x,y)}[f] \, dx \, dy
$$

$$
= \int_X p(x) \int_Y p(y|x) L_{(x,y)}[f] \, dy \, dx. \tag{19}
$$

Assuming that $\mathcal{L}$ is sufficiently smooth using the Euler-Lagrange derivative, we find that

$$
\nabla_f \mathcal{R}[f](x) = \frac{\partial (\mathcal{L}(x, f(x)))}{\partial f(x)}
$$

$$
= p(x) \int_Y p(y|x) \frac{\partial L_{(x,y)}[f]}{\partial f(x)} \, dy
$$

$$
= p(x) \cdot E_{y \sim p_y|x} \frac{\partial L_{(x,y)}[f]}{\partial f(x)}. \tag{20}
$$

A.1.2 FUNCTIONAL DERIVATIVE OF THE FUNCTIONAL ELBO

Let $F$ be an operator that depends on evaluation only with associated maps $F_X : \mathcal{Y}^s \rightarrow \mathcal{Y}^s, F_X : \mathcal{Y}^{s+M} \rightarrow \mathcal{Y}^{s+M}$. Further, let $\mathcal{F}[F] = \mathcal{L}_{D_s}(Q_{[T]}^{x}) = \frac{\mathbb{E}_{\tilde{y} \sim Q_{[T]}^{x}} \ell(D_s, \tilde{y} + F_{X}(\tilde{y}))}{\mathbb{D}_{KL}(Q_{[T]}^{x} || P_{f_X})}.

In general, and under the assumption that $\ell$ is sufficiently smooth, we find that

$$
\nabla_{F_X} \mathcal{F}_1[F_X] + \varepsilon G_{\tilde{y}} = \nabla_{F_X} \mathbb{E}_{\tilde{y} \sim Q_{X}} [\ell(D_s, \tilde{y} + F_{X}(\tilde{y})) + \varepsilon G_{X}(\tilde{y}) - \ell(D_s, \tilde{y} + F_{X}(\tilde{y}))] \tag{21}
$$

Under the assumption that $\mathcal{H}_{\mathcal{Y}} \subset \{f : \mathcal{Y}^s \rightarrow \mathcal{Y}^s\}$ is an RKHS with associated kernels $k_{\tilde{y}}$, we find that

$$
\nabla_{F_X} \mathcal{F}_1[F_X] = \mathbb{E}_{\tilde{y} \sim Q_{X}} [\nabla_{\tilde{y}} \ell(D_s, \tilde{y} + F_{X}(\tilde{y})) - \ell(D_s, \tilde{y} + F_{X}(\tilde{y}))]
$$

$$
= \varepsilon \mathbb{E}_{\tilde{y} \sim Q_{X}} [\nabla_{\tilde{y}} \ell(D_s, \tilde{y} + F_{X}(\tilde{y})) + O(\varepsilon^2)]
$$

$$
= \varepsilon \mathbb{E}_{\tilde{y} \sim Q_{X}} [\nabla_{\tilde{y}} \ell(D_s, \tilde{y} + F_{X}(\tilde{y})) + \nabla_{\tilde{y}} k_{\tilde{y}}(\tilde{y}, \cdot) + O(\varepsilon^2)] \tag{26}
$$

from which it follows that

$$
\nabla_{F_X} \mathcal{F}_1[F_X]|_{F_X=0} = \mathbb{E}_{\tilde{y} \sim Q_{X}} [\nabla_{\tilde{y}} \ell(D_s, \tilde{y} + F_{X}(\tilde{y})) + \nabla_{\tilde{y}} k_{\tilde{y}}(\tilde{y}, \cdot)]. \tag{29}
$$

Under the assumption that $\mathcal{H}_Y \subset \{f : \mathcal{Y}^{s+M} \rightarrow \mathcal{Y}^{s+M}\}$ is an RKHS with associated kernels $k_{\tilde{y}}$, it has been shown in [Liu & Wang, 2016] that

$$
\nabla_{F_X} \mathcal{F}_2[F_X]|_{F_X=0} = -\mathbb{E}_{\tilde{y} \sim Q_{X}} [\nabla_{\tilde{y}} \log p_{f_{\tilde{y}}}(y) k_{\tilde{y}}(y, \cdot) + \nabla_{\tilde{y}} k_{\tilde{y}}(y, \cdot)]. \tag{31}
$$

Using the chain rule, we obtain

$$
\nabla_{F} \mathcal{L}_{D_s,x}(Q_{[T]})|_{F=0}(f) = \nabla_{F_X} \mathcal{F}_1[F_X](f) \mathbb{E}_{\tilde{y}} \nabla_{F_X}(f) \bigg|_{F=0} - \nabla_{F_X} \mathcal{F}_2[F_X](f) \bigg|_{F=0}. \tag{33}
$$
We treat the weak learners. In its original form, tree-based learners were used as weak learners, which proved to be not scale the resulting sum of kernel functions. This is the natural learner if the hypothesis space \( H \) is assumed to be an RHKS, as discussed in section 2.1. The hyperparameters of the SFVGD step are the same as the ones we used for SFVNN. The results after 1000 iterations are summarized in Table 4.

### A.3 SFVGB

**A.3.1 SFVGB Algorithm**

We treat the \( r \) particle functions mapping from \( \mathcal{X} \) to \( \mathcal{Y} \) as a single function \( f[0] \) mapping from \( \mathcal{X} \) to \( \mathcal{Y}^r \) – i.e., we identify the \( i \)-th sample function \( f_i[0], i = 1, \ldots, r \) with the \( i \)-th component of \( f[0] \). Analogously to standard GB, we choose a base learner \( \mathcal{I}_b \) which defines a hypothesis space \( \mathcal{H} \subset \{ f : \mathcal{X} \rightarrow \mathcal{Y}^r \} \). While this requires the base learner \( \mathcal{I}_b \) to be a multi-output learner, it is always possible to use an ensemble of single-output learners. With this, SFVGB is vanilla GB of a multi-output function \( f \) where the loss is the negative functional ELBO. Consequently, we update \( f[t] \) in the \( t \)-th iteration via

\[
\begin{align*}
 f[t+1] &= f[t] + \eta[t] \mathcal{I}_b \left( \mathbf{X}, \left[ \nabla_F \mathcal{L}_{D_s, \mathbf{X}(Q[T])} \big|_{F=0} (f[t]) (\mathbf{X}), \ldots, \nabla_F \mathcal{L}_{D_s, \mathbf{X}(Q[T])} \big|_{F=0} (f[t]) (\mathbf{X}) \right] \right),
\end{align*}
\]

using Eq. 18 and a (potentially adaptive) learning rate \( \eta[t] \).

**A.3.2 SFVGB Regression Experiments**

We also test SFVGB on the small regression datasets and compare it to 1) the approach proposed in Malinin et al. (2021), i.e., uncertainty quantification approaches by randomly subsampling the data in every iteration of a stochastic gradient boosting (GB) model and 2) boosting generalized linear model (GLMB Buehlmaj 2006) as a baseline approach. As base learners, SGB uses trees and GLMB uses linear models. Our model uses a Nadaraya-Watson kernel regression variant, which does not scale the resulting sum of kernel functions. This is the natural learner if the hypothesis space \( \mathcal{H} \) is assumed to be an RHKS, as discussed in section 2.1. The hyperparameters of the SFVGD step are the same as the ones we used for SFVNN. The results after 1000 iterations are summarized in Table 4.

---

**Algorithm 3:** Stein Functional Variational Gradient Boosting

**Hyperparameters:** Same as for sfvgd_step

**Input:** number of iterations \( t_{\text{max}} \), learning rate \( \eta[t] \) in the \( t \)-th iteration, set of initial particle functions \( \{ f_i[0] \}_{i=1} \) treated as multi-output function \( f[0] \), multi-output base learner \( \mathcal{I}_b \)

**Output:** Set of particle functions \( \{ f_i[t_{\text{max}}] \}_{i=1} \), which approximate the target distribution

for \( t = 0, \ldots, t_{\text{max}} - 1 \) do

\[
\begin{align*}
 \mathbf{X}_t, \Delta_{\mathbf{X}_t} &= \text{sfvgd_step}(f[t]) \\
 f[t+1] &= f[t] + \eta[t] \cdot \mathcal{I}_b(\mathbf{X}, \Delta_{\mathbf{X}_t})
\end{align*}
\]

end
Table 3: Comparison of boosting approaches on the small benchmark data sets (columns) using the average NLL (smaller is better) over 10 train-test data splits with standard deviation in brackets. The best performing method for each data set is highlighted in bold.

<table>
<thead>
<tr>
<th></th>
<th>Airfoil</th>
<th>Concrete</th>
<th>Diabetes</th>
<th>Energy</th>
<th>ForestF</th>
<th>Wine</th>
<th>Yacht</th>
</tr>
</thead>
<tbody>
<tr>
<td>SFVGB</td>
<td>2.87 (0.18)</td>
<td>3.41 (0.13)</td>
<td>7.04 (0.67)</td>
<td>2.16 (0.10)</td>
<td>6.31 (0.69)</td>
<td>2.37 (2.00)</td>
<td>3.21 (0.52)</td>
</tr>
<tr>
<td>GLMB</td>
<td>3.10 (0.03)</td>
<td>3.93 (0.03)</td>
<td>5.50 (0.03)</td>
<td>3.28 (0.01)</td>
<td>1.84 (0.07)</td>
<td>0.72 (0.02)</td>
<td>3.80 (0.04)</td>
</tr>
<tr>
<td>SGB</td>
<td>1.98 (0.05)</td>
<td>3.06 (0.10)</td>
<td>5.51 (0.07)</td>
<td>0.79 (0.49)</td>
<td>1.86 (0.08)</td>
<td>0.11 (0.44)</td>
<td>0.36 (0.23)</td>
</tr>
</tbody>
</table>

Table 4: Comparison of boosting approaches on the small benchmark data sets (columns) using the average RMSE (smaller is better) over 10 train-test data splits with standard deviation in brackets. The best performing method for each data set is highlighted in bold.

<table>
<thead>
<tr>
<th></th>
<th>Airfoil</th>
<th>Concrete</th>
<th>Diabetes</th>
<th>Energy</th>
<th>ForestF</th>
<th>Wine</th>
<th>Yacht</th>
</tr>
</thead>
<tbody>
<tr>
<td>SFVGB</td>
<td>3.31 (0.26)</td>
<td>6.71 (0.52)</td>
<td>57.8 (0.67)</td>
<td>1.93 (0.14)</td>
<td>1.53 (0.07)</td>
<td>0.18 (0.04)</td>
<td>4.83 (1.24)</td>
</tr>
<tr>
<td>GLMB</td>
<td>4.85 (0.27)</td>
<td>10.5 (1.03)</td>
<td>53.3 (3.42)</td>
<td>3.07 (0.21)</td>
<td>1.51 (0.09)</td>
<td>0.27 (0.04)</td>
<td>8.51 (1.18)</td>
</tr>
<tr>
<td>SGB</td>
<td>2.06 (0.21)</td>
<td>5.06 (0.53)</td>
<td>57.6 (2.71)</td>
<td>0.57 (0.09)</td>
<td>1.52 (0.09)</td>
<td>0.29 (0.14)</td>
<td>0.85 (0.45)</td>
</tr>
</tbody>
</table>

Further experimental details  Model tuning of SGB is done as explained in Malinin et al. (2021). Here, different tree depths ∈ {3, 4, 5, 6}, learning rates {0.001, 0.01, 0.1}, and numbers of samples ∈ {0.25, 0.5, 0.75} of approaches are trained on the first 80% of the training data and evaluated on the latter 20%. The GLMB approach is tuned using a 10-fold cross-validation to determine the number of stopping iterations, which is the only hyperparameter of the model.

Results  Results show that the SFVGB can often improve over the GLMB baseline but still yields inferior performance on some data sets. This is, in particular, the case if there is a rather discrete outcome space (e.g., Wine which only consists of values 0, 1, and 2). The SGB model, in turn, works better than both the SFVGB and GLMB in most cases. This is likely due to the much more flexible base learner structure (as SGB like most of the state-of-the-art boosting approaches uses trees, whereas GLMB uses linear regression and ours uses kernel regression).

A.4 Numerical Experiments: Further Details

A.4.1 Further Data Details

We use the benchmark data setup proposed by Hernandez-Lobato & Adams (2015) and Sun et al. (2019) for evaluating probabilistic regression approaches. This setup includes selected data sets from the UCI repository – namely, the four smaller data sets Concrete, Energy, Wine, Yacht, and the four larger data sets Naval, Protein, Video (Memory and Time), and GPU. In addition to Sun et al. (2019), we also compare our approaches on three additional smaller data sets (Airfoil, Diabetes, Forest Fire) and further investigate the second task on the Naval data set (i.e., we examine both the compressor decay and the turbine decay state coefficient, referred to as NavalC and NavalT, respectively). In Table 5, the data characteristics and pre-processing steps are listed.

A.4.2 Further Experimental Details

BNN approaches are fitted using the recommended architecture and tuning parameters by Sun et al. (2019). We reduced the epochs from 10,000 to 1,000 epochs for the smaller data sets to reduce the computational runtime. This did not negatively impact the BNN’s performance. In all our benchmark experiments, we follow the setup for BNNs and FVBNns described by Sun et al. (2019).

A.5 Further Results

Here, we provide further results using the application of probabilistic methods for contextual bandits.

Contextual Bandits  One important application of uncertainty-aware models is for exploration, as in Bayesian optimization, reinforcement learning, or bandits. Following Sun et al. (2019), we evaluate SFVFNN using the contextual bandits benchmark by Riquelme et al. (2018) by re-running the settings investigated in Sun et al. (2019) and report the cumulative regret based on the best expected
Table 5: Data set characteristics, additional pre-processing and references.

<table>
<thead>
<tr>
<th>Dataset</th>
<th># Obs.</th>
<th># Feat.</th>
<th>Pre-processing</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Airfoil</td>
<td>1503</td>
<td>5</td>
<td>-</td>
<td>Dua &amp; Graff (2017)</td>
</tr>
<tr>
<td>Concrete</td>
<td>1030</td>
<td>8</td>
<td>-</td>
<td>Yeh (1998)</td>
</tr>
<tr>
<td>Diabetes</td>
<td>442</td>
<td>10</td>
<td>-</td>
<td>Dua &amp; Graff (2017)</td>
</tr>
<tr>
<td>Energy</td>
<td>768</td>
<td>8</td>
<td>-</td>
<td>Tsanas &amp; Xifara (2012)</td>
</tr>
<tr>
<td>ForestF</td>
<td>517</td>
<td>12</td>
<td>logp1 transformation for area; numerical representation for month and day</td>
<td>Cortez &amp; Morais (2007)</td>
</tr>
<tr>
<td>Wine</td>
<td>178</td>
<td>13</td>
<td>-</td>
<td>Dua &amp; Graff (2017)</td>
</tr>
<tr>
<td>Yacht</td>
<td>308</td>
<td>6</td>
<td>-</td>
<td>Dua &amp; Graff (2017)</td>
</tr>
<tr>
<td>GPU</td>
<td>241600</td>
<td>14</td>
<td>only use run 1 as outcome</td>
<td>Ballester-Ripoll et al. (2017)</td>
</tr>
<tr>
<td>NavalT</td>
<td>11934</td>
<td>15</td>
<td>drop features with zero variance</td>
<td>Coraddu et al. (2014)</td>
</tr>
<tr>
<td>NavalC</td>
<td>11934</td>
<td>15</td>
<td>drop features with zero variance</td>
<td>Coraddu et al. (2014)</td>
</tr>
<tr>
<td>Protein</td>
<td>45730</td>
<td>9</td>
<td>-</td>
<td>Dua &amp; Graff (2017)</td>
</tr>
<tr>
<td>Video</td>
<td>68784</td>
<td>19</td>
<td>drop highly correlated features; drop id and b_size; use dummy-coding for codec and o_codec</td>
<td>Dua &amp; Graff (2017)</td>
</tr>
</tbody>
</table>

Table 6: Relative contextual bandits regret (relative to the cumulative regret of Uniform sampling) for different data sets (columns) and methods (rows). Numbers in brackets of methods indicate the network sizes. Reported numbers are the mean (and standard derivation in brackets) over 5 trials. The best algorithms per data set are highlighted in bold.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>BNN (50)</th>
<th>BNN (500)</th>
<th>BootRMS</th>
<th>Dropout</th>
<th>FVBNN (50, 50, 50)</th>
<th>FVBNN (50)</th>
<th>ParamNoise</th>
<th>SFVNN (50, 50, 50)</th>
<th>SFVNN (50)</th>
<th>Uniform Sampling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adult</td>
<td>95.79 (1.33)</td>
<td>99.16 (1.47)</td>
<td>82.74 (8.16)</td>
<td>85.46 (4.74)</td>
<td>72.84 (11.5)</td>
<td>75.05 (7.66)</td>
<td>89.00 (5.24)</td>
<td>71.99 (7.29)</td>
<td>79.62 (5.44)</td>
<td>100.0 (0.00)</td>
</tr>
<tr>
<td>Census</td>
<td>66.12 (5.10)</td>
<td>94.38 (11.9)</td>
<td>53.40 (18.0)</td>
<td>37.27 (12.3)</td>
<td>30.46 (28.4)</td>
<td>40.65 (14.9)</td>
<td>53.86 (15.5)</td>
<td>36.65 (30.3)</td>
<td>30.05 (13.1)</td>
<td>100.0 (0.00)</td>
</tr>
<tr>
<td>Covertype</td>
<td>61.18 (1.91)</td>
<td>74.55 (7.32)</td>
<td>36.95 (10.3)</td>
<td>37.27 (12.3)</td>
<td>24.29 (14.7)</td>
<td>46.57 (3.99)</td>
<td>48.18 (9.84)</td>
<td>28.47 (30.3)</td>
<td>29.24 (11.4)</td>
<td>100.0 (0.00)</td>
</tr>
<tr>
<td>Jester</td>
<td>83.70 (3.31)</td>
<td>79.08 (5.84)</td>
<td>63.3 (10.74)</td>
<td>65.51 (9.62)</td>
<td>50.45 (20.0)</td>
<td>57.28 (17.8)</td>
<td>66.52 (13.1)</td>
<td>50.09 (21.2)</td>
<td>29.24 (11.4)</td>
<td>100.0 (0.00)</td>
</tr>
<tr>
<td>Mushroom</td>
<td>7.81 (9.39)</td>
<td>9.31 (11.1)</td>
<td>7.11 (5.61)</td>
<td>4.43 (6.83)</td>
<td>2.92 (10.3)</td>
<td>2.39 (6.24)</td>
<td>6.74 (5.89)</td>
<td>7.81 (9.35)</td>
<td>29.24 (11.4)</td>
<td>100.0 (0.00)</td>
</tr>
<tr>
<td>Statlog</td>
<td>29.2 (4.05)</td>
<td>64.5 (13.3)</td>
<td>1.81 (2.70)</td>
<td>2.89 (4.26)</td>
<td>3.46 (3.75)</td>
<td>1.58 (2.09)</td>
<td>7.69 (3.57)</td>
<td>4.52 (4.12)</td>
<td>91.95 (33.2)</td>
<td>100.0 (0.00)</td>
</tr>
<tr>
<td>Wheel</td>
<td>82.94 (22.2)</td>
<td>8.892 (15.0)</td>
<td>119.16 (3.7)</td>
<td>28.61 (12.1)</td>
<td>13.62 (14.7)</td>
<td>24.24 (22.0)</td>
<td>21.93 (15.4)</td>
<td>44.12 (27.7)</td>
<td>91.95 (33.2)</td>
<td>100.0 (0.00)</td>
</tr>
</tbody>
</table>

reward. Following Riquelme et al. (2018), we use the benchmark data sets Adult, Census, Covertype, Financial, Jester, Mushroom, Statlog, and Wheel. For these, the rewards are deterministic, and the regret is equal to the best realized reward. As in previous works, we report the regret as a relative value, relative to a random uniform sampling procedure that emulates Thompson Sampling (see Riquelme et al., 2018). As comparison methods, we use the BNN (with 50 and 500 units), three spinnoffs of the NeuralLinear algorithm (namely, the Bootstrapped NN trained with RMSprop (BootRMS), Parameter Noise (ParamNoise), and Dropout (see Riquelme et al., 2018, for more details)), as well as two variants of the FVBNN (with one and three layers each with 50 units). Experiments are run 5 times with shuffled contexts, for which we report mean and standard deviation of the relative cumulative regret.

Results are given in Table 6 suggesting that both FVBNN and SFVNN are well- and particularly similar-performing methods in the application of contextual bandits.

B RUNTIME EXPERIMENT

B.1 SETUP

We trained FVBN and SFVNN 5 times on the Energy data set for each number of particle functions \( r \in \{50, 100, 150\} \) and plotted the resulting runtimes in Figure 2. It becomes apparent that the number of particle functions influences the runtime of FVBN more substantially than SFVNN, as we expect from our computational complexity analysis.
Figure 2: Comparison of the runtimes of FVBNN and SFVNN on the Energy data set with 5 repetitions for each number of particle functions

B.2 COMPUTATIONAL ENVIRONMENT

All experiments and benchmarks were carried out on an internal cluster with Intel® Xeon® CPU E5-2650 v2 @ 2.60GHz, 32 cores, 64 GB Random-access memory, and operating system Ubuntu 20.04.1 LTS.