

# A Framework and Benchmark for Deep Batch Active Learning for Regression

**David Holzmüller**

DAVID.HOLZMUELLER@MATHEMATIK.UNI-STUTTGART.DE

*University of Stuttgart  
Faculty of Mathematics and Physics  
Institute for Stochastics and Applications*

**Viktor Zaverkin\***

ZAVERKIN@THEOCHEM.UNI-STUTTGART.DE

*University of Stuttgart  
Faculty of Chemistry  
Institute for Theoretical Chemistry*

**Johannes Kästner**

KAESTNER@THEOCHEM.UNI-STUTTGART.DE

*University of Stuttgart  
Faculty of Chemistry  
Institute for Theoretical Chemistry*

**Ingo Steinwart**

INGO.STEINWART@MATHEMATIK.UNI-STUTTGART.DE

*University of Stuttgart  
Faculty of Mathematics and Physics  
Institute for Stochastics and Applications*

**Editor:** Aarti Singh

## Abstract

The acquisition of labels for supervised learning can be expensive. To improve the sample efficiency of neural network regression, we study active learning methods that adaptively select batches of unlabeled data for labeling. We present a framework for constructing such methods out of (network-dependent) base kernels, kernel transformations, and selection methods. Our framework encompasses many existing Bayesian methods based on Gaussian process approximations of neural networks as well as non-Bayesian methods. Additionally, we propose to replace the commonly used last-layer features with sketched finite-width neural tangent kernels and to combine them with a novel clustering method. To evaluate different methods, we introduce an open-source benchmark consisting of 15 large tabular regression data sets. Our proposed method outperforms the state-of-the-art on our benchmark, scales to large data sets, and works out-of-the-box without adjusting the network architecture or training code. We provide open-source code that includes efficient implementations of all kernels, kernel transformations, and selection methods, and can be used for reproducing our results.

**Keywords:** batch mode deep active learning, regression, tabular data, neural network, benchmark

---

\*. Present address: NEC Laboratories Europe GmbH, Kurfürsten-Anlage 36, 69115 Heidelberg, Germany

## 1. Introduction

While supervised machine learning (ML) has been successfully applied to many different problems, these successes often rely on the availability of large data sets for the problem at hand. In cases where labeling data is expensive, it is important to reduce the required number of labels. Such a reduction could be achieved through various means: First, finding more sample-efficient supervised ML methods; second, applying data augmentation; third, leveraging information in unlabeled data via semi-supervised learning; fourth, leveraging information from related problems through transfer learning, meta-learning, or multi-task learning; and finally, appropriately selecting which data to label. Active learning (AL) takes the latter approach by using a trained model to choose the next data point to label (Settles, 2009). The need to retrain after every new label prohibits parallelized labeling methods and can be far too expensive, especially for neural networks (NNs), which are often slow to train. This problem can be resolved by batch mode active learning (BMAL) methods, which select multiple data points for labeling at once. When the supervised ML method is a deep NN, this is known as batch mode deep active learning (BMDAL) (Ren et al., 2021). Pool-based BMDAL refers to the setting where data points for labeling need to be chosen from a given finite set of points.

Supervised and unsupervised ML algorithms choose a model for given data. Multiple models can be compared on the same data using model selection techniques such as cross-validation. Such a comparison increases the training cost, but not the (potentially much larger) cost of labeling data. In contrast to supervised learning, AL is about choosing the data itself, with the goal to reduce labeling cost. However, different AL algorithms may choose different samples, and hence a comparison of  $N$  AL algorithms might increase labeling cost by a factor of up to  $N$ . Consequently, such a comparison is not sensible for applications where labeling is expensive. Instead, it is even more important to properly benchmark AL methods on tasks where labels are cheap to generate or a large number of labels is already available.

In the classification setting, NNs typically output uncertainties in the form of a vector of probabilities obtained through a softmax layer, while regression NNs typically output a scalar target without uncertainties. Therefore, many BMDAL algorithms only apply to one of the two settings. For classification, many BMDAL approaches have been proposed (Ren et al., 2021), and there exist at least some standard benchmark data sets like CIFAR-10 (Krizhevsky, 2009) on which methods are usually evaluated. On the other hand, the regression setting has been studied less frequently, and no common benchmark has been established to the best of our knowledge, except for a specialized benchmark in drug discovery (Mehrjou et al., 2021). We expect that the regression setting will gain popularity, not least due to the increasing interest in NNs for surrogate modeling (Behler, 2016; Kutz, 2017; Raissi et al., 2019; Mehrjou et al., 2021; Lavin et al., 2021).

### 1.1 Contributions

In this paper, we investigate pool-based BMDAL methods for regression. Our experiments use fully connected NNs on tabular data sets, but the considered methods can be generalized to different types of data and NN architectures. We limit our study to methods that do not require to modify the network architecture and training, as these are particularly easy to

use and a fair comparison to other methods is difficult. We also focus on methods that scale to large amounts of (unlabeled) data and large acquisition batch sizes. Our contributions can be summarized as follows:

- (1) We propose a framework for decomposing typical BM(D)AL algorithms into the choice of a kernel and a selection method. Here, the kernel can be constructed from a base kernel through a series of kernel transformations. The use of kernels as basic building blocks allows for an efficient yet flexible and composable implementation of our framework, which we include in our open-source code. We also discuss how (regression variants of) many popular BM(D)AL algorithms can be represented in this framework and how they can efficiently be implemented. This gives us a variety of options for base kernels, kernel transformations, and selection methods to combine. Our framework encompasses both Bayesian methods based on Gaussian Processes and Laplace approximations as well as geometric methods.
- (2) We discuss some alternative options to the ones arising from popular BM(D)AL algorithms: We introduce a novel selection method called LCMD; and we propose to combine the finite-width neural tangent kernel (NTK, Jacot et al., 2018) as a base kernel with sketching for efficient computation.
- (3) We introduce an open-source benchmark for BMDAL involving 15 large tabular regression data sets. Using this benchmark, we compare different selection methods and evaluate the influence of the kernel, the acquisition batch size, and the target metric.

Our newly proposed selection method, LCMD, improves the state-of-the-art in our benchmark in terms of RMSE and MAE, while still exhibiting good performance for the maximum error. The NTK base kernel improves the benchmark accuracy for all selection methods, and the proposed sketching method can preserve this accuracy while leading to significant time gains. Figure 1 shows a comparison of our novel BMDAL algorithm against popular BMDAL algorithms from the literature, which are all implemented in our framework. The code for our framework and benchmark is based on PyTorch (Paszke et al., 2019) and is publicly available at

[https://github.com/dholzmueLLer/bmdal\\_reg](https://github.com/dholzmueLLer/bmdal_reg)

and will be archived together with the generated data at <https://doi.org/10.18419/darus-3394>.

The rest of this paper is structured as follows: In Section 2, we introduce the basic problem setting of BMDAL for tabular regression with fully-connected NNs and introduce our framework for the construction of BMDAL algorithms. We discuss related work in Section 3. We then introduce options to build kernels from base kernels and kernel transformations in Section 4. Section 5 discusses various iterative kernel-based selection methods. Our experiments in Section 6 provide insights into the performance of different combinations of kernels and selection methods. Finally, we discuss limitations and open questions in Section 7. More details on the presented methods and experimental results are provided in the Appendix, whose structure is outlined in Section A.

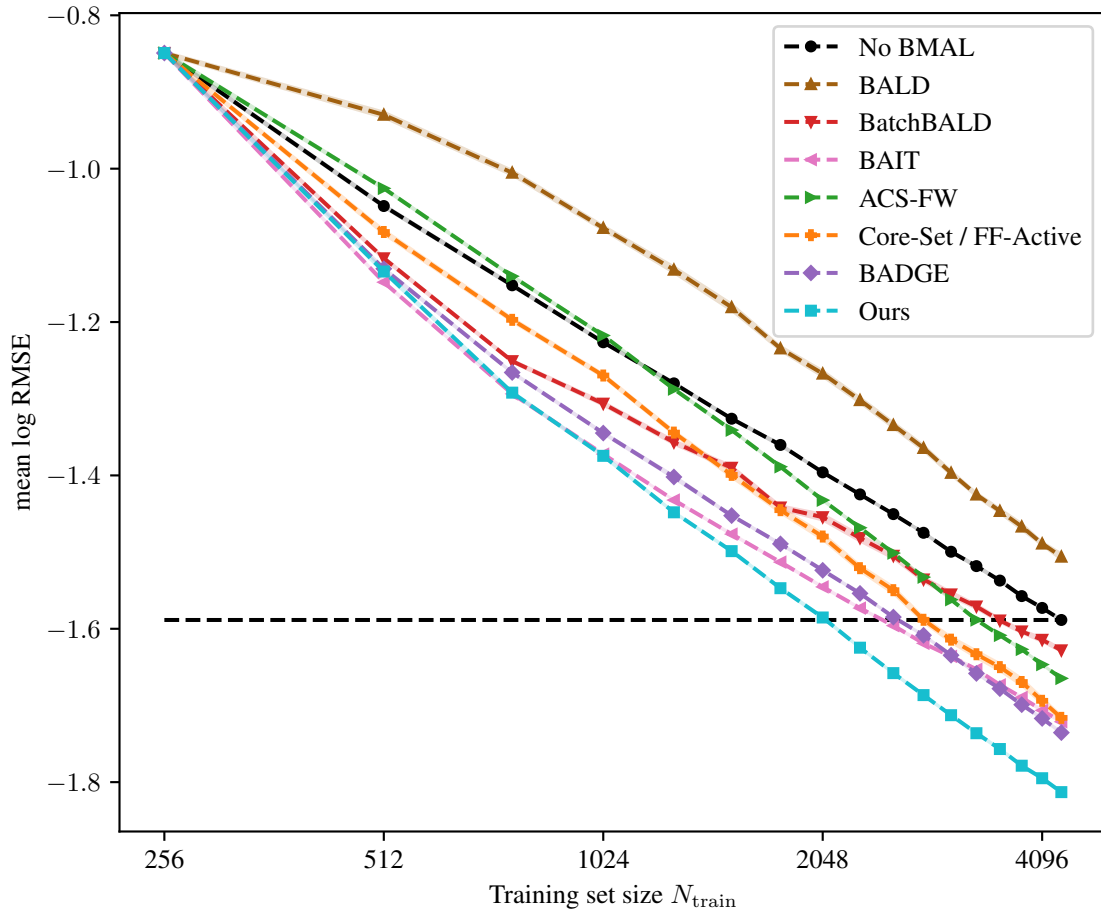


Figure 1: This figure shows how fast the averaged errors on our benchmark data sets decrease during BMAL for random selection (no BMAL), BALD (Houlsby et al., 2011), BatchBALD (Kirsch et al., 2019), BAIT (Ash et al., 2021), ACS-FW (Pinsler et al., 2019), Core-Set (Sener and Savarese, 2018), FF-Active (Geifman and El-Yaniv, 2017), BADGE (Ash et al., 2019), and our method. In Table 5 and Section 6, we specify how the compared methods are built from components explained in Section 4 and Section 5, and discuss further details such as modifications to apply them to regression. For the plot, we start with 256 random training samples and select 256 samples in each of 16 BMAL steps. The lines show the average of the logarithmic RMSE over all 15 benchmark data sets and 20 random splits between the BMAL steps. The shaded area, which is barely visible, corresponds to one estimated standard deviation of the mean estimator, cf. Section E.4.

## 2. Problem Setting

In this section, we outline the problem of BMDAL for regression with fully-connected NNs. We first introduce the regression objective and fully-connected NNs. Subsequently, we introduce the basic setup of pool-based BMDAL as well as our proposed framework.

## 2.1 Regression with Fully-Connected Neural Networks

We consider multivariate regression, where the goal is to learn a function  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  from data  $(\mathbf{x}, y) \in \mathcal{D}_{\text{train}} \subseteq \mathbb{R}^d \times \mathbb{R}$ . In the case of NNs, we consider a parameterized function family  $(f_{\boldsymbol{\theta}})_{\boldsymbol{\theta} \in \mathbb{R}^m}$  and try to minimize the mean squared loss on training data  $\mathcal{D}_{\text{train}}$  with  $N_{\text{train}}$  samples:

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{N_{\text{train}}} \sum_{(\mathbf{x}, y) \in \mathcal{D}_{\text{train}}} (y - f_{\boldsymbol{\theta}}(\mathbf{x}))^2 .$$

We refer to the inputs and labels in  $\mathcal{D}_{\text{train}}$  as  $\mathcal{X}_{\text{train}}$  and  $\mathcal{Y}_{\text{train}}$ , respectively. Corresponding data sets are often referred to as *tabular data* or *structured data*. This is in contrast to data with a known spatiotemporal relationship between the input features, such as image or time-series data, where specialized NN architectures such as CNNs are more successful.

For our derivations and experiments, we consider an  $L$ -layer fully-connected NN  $f_{\boldsymbol{\theta}} : \mathbb{R}^d \rightarrow \mathbb{R}$  with parameter vector  $\boldsymbol{\theta} = (\mathbf{W}^{(1)}, \mathbf{b}^{(1)}, \dots, \mathbf{W}^{(L)}, \mathbf{b}^{(L)})$  and input size  $d_0 = d$ , hidden layer sizes  $d_1, \dots, d_{L-1}$ , and output size  $d_L = 1$ . The value  $\mathbf{z}_i^{(L)} = f_{\boldsymbol{\theta}}(\mathbf{x}_i^{(0)})$  of the NN on the  $i$ -th input  $\mathbf{x}_i^{(0)} \in \mathbb{R}^{d_0}$  is defined recursively by

$$\mathbf{x}_i^{(l+1)} = \varphi(\mathbf{z}_i^{(l+1)}) \in \mathbb{R}^{d_{l+1}}, \quad \mathbf{z}_i^{(l+1)} = \frac{\sigma_w}{\sqrt{d_l}} \mathbf{W}^{(l+1)} \mathbf{x}_i^{(l)} + \sigma_b \mathbf{b}^{(l+1)} \in \mathbb{R}^{d_{l+1}} . \quad (1)$$

Here, the activation function  $\varphi : \mathbb{R} \rightarrow \mathbb{R}$  is applied element-wise and  $\sigma_w, \sigma_b > 0$  are constant factors. In our experiments, the weight matrices are initialized with independent standard normal entries and the biases are initialized to zero. The factors  $\sigma_w/\sqrt{d_l}$  and  $\sigma_b$  stem from the neural tangent parametrization (NTP) (Jacot et al., 2018; Lee et al., 2019), which is theoretically motivated to define infinite-width limits of NNs and is also used in our applications. However, our derivations apply analogously to NNs without these factors. When considering different NN types such as CNNs, it is possible to apply our derivations only to the fully-connected part of the NN or to extend them to other layers as well.

## 2.2 Batch Mode Active Learning

In a single BMAL step, a BMAL algorithm selects a batch  $\mathcal{X}_{\text{batch}} \subseteq \mathbb{R}^d$  with a given size  $N_{\text{batch}} \in \mathbb{N}$ . Subsequently, this batch is labeled and added to the training set. Here, we consider *pool-based* BMAL, where  $\mathcal{X}_{\text{batch}}$  is to be selected from a given finite *pool set*  $\mathcal{X}_{\text{pool}}$  of candidates. Other AL paradigms include membership query AL, where data points for labeling can be chosen freely, or stream-based AL, where data points arrive sequentially and must be immediately labeled or discarded. The pool set can potentially contain information about which regions of the input space are more important than others, especially if it is drawn from the same distribution as the test set. Moreover, pool-based BMAL allows for efficient benchmarking of BMAL methods on labeled data sets by reserving a large portion of the data set for the pool set, rendering the labeling part trivial.

When comparing and evaluating BMDAL methods, we are mainly interested in the following desirable properties:

- (P1) The method should improve the sample efficiency of the underlying NN, even for large acquisition batch sizes  $N_{\text{batch}}$  and large pool set sizes  $N_{\text{pool}}$ , with respect to the

---

**Algorithm 1** Basic pool-based BMDAL loop with initial labeled training set  $\mathcal{D}_{\text{train}}$ , unlabeled pool set  $\mathcal{X}_{\text{pool}}$ , BMDAL algorithm NEXTBATCH (see Algorithm 2) and a list  $L_{\text{batch}}$  of batch sizes.

---

```

for AL batch size  $N_{\text{batch}}$  in  $L_{\text{batch}}$  do
    Train model  $f_{\theta}$  on  $\mathcal{D}_{\text{train}}$ 
    Select batch  $\mathcal{X}_{\text{batch}} \leftarrow \text{NEXTBATCH}(f_{\theta}, \mathcal{D}_{\text{train}}, \mathcal{X}_{\text{pool}}, N_{\text{batch}})$  with  $|\mathcal{X}_{\text{batch}}| = N_{\text{batch}}$ 
    and  $\mathcal{X}_{\text{batch}} \subseteq \mathcal{X}_{\text{pool}}$ 
    Move  $\mathcal{X}_{\text{batch}}$  from  $\mathcal{X}_{\text{pool}}$  to  $\mathcal{D}_{\text{train}}$  and acquire labels  $\mathcal{Y}_{\text{batch}}$  for  $\mathcal{X}_{\text{batch}}$ 
end for
Train final model  $f_{\theta}$  on  $\mathcal{D}_{\text{train}}$ 

```

---

downstream application, which may or may not involve the same input distribution as training and pool data.

- (P2) The method should scale to large pool sets, training sets, and batch sizes, in terms of both computation time and memory consumption.
- (P3) The method should be applicable to a wide variety of NN architectures and training methods, such that it can be applied to different use cases.
- (P4) The method should not require modifying the NN architecture and training method, for example by requiring to introduce Dropout, such that practitioners do not have to worry whether employing the method diminishes the accuracy of their trained NN.
- (P5) The method should not require training multiple NNs for a single batch selection since this would deteriorate its runtime efficiency.<sup>1</sup>
- (P6) The method should not require tuning hyperparameters on the downstream application since this would require labeling samples selected with suboptimal hyperparameters.

Property (P1) is central to motivate the use of BMDAL over random sampling of the data and is evaluated for our framework in detail in Section 6 and Section E. We only evaluate methods with property (P2) since our benchmark involves large data sets. All methods considered here satisfy (P3) to a large extent. Indeed, although efficient computations are only studied for fully-connected layers here, the considered methods can be simply applied to the fully-connected part of a larger NN. All considered methods satisfy (P4), which also facilitates fair comparison in a benchmark. All our methods satisfy (P5), although our methods can incorporate ensembles of NNs. Although some of the considered methods have hyperparameters, we fix them to reasonable values independent of the data set in our experiments, such that (P6) is satisfied.

Algorithm 1 shows how BMDAL algorithms satisfying (P4) and (P5) can be used in a loop with training and labeling.

Wu (2018) formulates three criteria by which BMAL algorithms may select batch samples in order to improve the sample efficiency of a learning method:

- (INF) The algorithm should favor inputs that are *informative* to the model. These could, for example, be those inputs where the model is most uncertain about the label.
- (DIV) The algorithm should ensure that the batch contains *diverse* samples, i.e., samples in the batch should be sufficiently different from each other.

---

1. Technically, requiring multiple trained NNs would not be detrimental if it facilitated reaching the same accuracy with correspondingly larger  $N_{\text{batch}}$ .

---

**Algorithm 2** Kernel-based batch construction framework

---

```

function KERNELNEXTBATCH( $f_{\theta}$ ,  $\mathcal{D}_{\text{train}}$ ,  $\mathcal{X}_{\text{pool}}$ ,  $N_{\text{batch}}$ )
   $k \leftarrow$  BASEKERNEL( $f_{\theta}$ )
   $k \leftarrow$  TRANSFORMKERNEL( $k$ ,  $\mathcal{D}_{\text{train}}$ )
  return SELECT( $k$ ,  $\mathcal{X}_{\text{train}}$ ,  $\mathcal{X}_{\text{pool}}$ ,  $N_{\text{batch}}$ )
end function

```

---

(REP) The algorithm should ensure *representativity* of the resulting training set, i.e., it should focus more strongly on regions where the pool data distribution has high density.

Note that (REP) might not be desirable if one expects a significant distribution shift between pool and test data. A challenge in trying to adapt non-batch AL methods to the batch setting is that some non-batch AL methods expect to immediately receive a label for every selected sample. It is usually possible to circumvent this by selecting the  $N_{\text{batch}}$  samples with the largest acquisition function scores at once, but this does not enforce (DIV) or (REP).

We propose a framework for assembling BMDAL algorithms that is shown in Algorithm 2 and consists of three components: First, a base kernel  $k$  needs to be chosen that should serve as a proxy for the trained network  $f_{\theta}$ . Second, the kernel can be transformed using various transformations. These transformations can, for example, make the kernel represent posteriors or improve its evaluation efficiency. Third, a selection method is invoked that uses the transformed kernel as a measure of similarity between inputs. When using Gaussian Process regression with a given kernel  $k$  as a supervised learning method instead of an NN, the base kernel could simply be chosen as  $k$ . Note that SELECT does not observe the training labels directly, however, in the NN setting, these can be implicitly incorporated through kernels that depend on the trained NN.

**Example 1.** *In Algorithm 2, the base kernel  $k$  could be of the form  $k(\mathbf{x}, \tilde{\mathbf{x}}) = \langle \phi(\mathbf{x}), \phi(\tilde{\mathbf{x}}) \rangle$ , where  $\phi$  represents the trained NN without the last layer. When interpreting  $k$  as the kernel of a Gaussian process, TRANSFORMKERNEL could then compute a transformed kernel  $\tilde{k}$  that represents the posterior predictive uncertainty after observing the training data. Finally, SELECT could then choose the  $N_{\text{batch}}$  points  $\mathbf{x} \in \mathcal{X}_{\text{pool}}$  with the largest uncertainty  $\tilde{k}(\mathbf{x}, \mathbf{x})$ .*

From a Bayesian perspective, our choice of kernel and kernel transformations can correspond to inference in a Bayesian approximation, as we discuss in Section C.1, while the selection method can correspond to the optimization of an acquisition function. However, in our framework, the same ‘‘Bayesian’’ kernels can be used together with non-Bayesian selection methods and vice versa.

### 3. Related Work

The field of active learning, also known as query learning or sequential (optimal) experimental design (Fedorov, 1972; Chaloner and Verdinelli, 1995), has a long history dating back at least to the beginning of the 20th century (Smith, 1918). For an overview of the AL and BMDAL literature, we refer to Settles (2009); Kumar and Gupta (2020); Ren et al. (2021); Weng (2022).

We first review work relevant to the kernels in our framework, before discussing work more relevant to selection methods, and finally, data sets. More literature related to specific methods is also discussed in Section 4 and Section 5.

### 3.1 Uncertainty Measures and Kernel Approximations

A popular class of BMDAL methods is given by Bayesian methods since the Bayesian framework naturally provides uncertainties that can be used to assess informativeness. These methods require to use Bayesian NNs, or in other words, the calculation of an approximate posterior distribution over NN parameters. A simple option is to perform Bayesian inference only over the last layer of the NN (Lázaro-Gredilla and Figueiras-Vidal, 2010; Snoek et al., 2015; Ober and Rasmussen, 2019; Kristiadi et al., 2020). The Laplace approximation (Laplace, 1774; MacKay, 1992a) can provide a local posterior distribution around a local optimum of the loss landscape via a second-order Taylor approximation. An alternative local approach based on SGD iterates is called SWAG (Maddox et al., 2019). Ensembles of NNs (Hansen and Salamon, 1990; Lakshminarayanan et al., 2017) can be interpreted as a simple multi-modal posterior approximation and can be combined with local approximations to yield mixtures of Laplace approximations (Eschenhagen et al., 2021) or MultiSWAG (Wilson and Izmailov, 2020). Monte Carlo (MC) Dropout (Gal and Ghahramani, 2016) is an option to obtain ensemble predictions from a single NN, although it requires training with Dropout (Srivastava et al., 2014). Regarding uncertainty approximations, our considered algorithms are mainly related to exact last-layer methods and the Laplace approximation, as these do not require to modify the training process. Daxberger et al. (2021) give an overview of various methods to compute (approximate) Laplace approximations.

Some recent approaches also build on the neural tangent kernel (NTK) introduced by Jacot et al. (2018). Khan et al. (2019) show that certain Laplace approximations are related to the finite-width NTK. Wang et al. (2022) and Mohamadi et al. (2022) propose the use of finite-width NTKs for DAL for classification. Wang et al. (2021) use the finite-width NTK at initialization for the streaming setting of DAL for classification and theoretically analyze the resulting method. Aljundi et al. (2022) use a kernel related to the finite-width NTK for DAL. Shoham and Avron (2020), Borsos et al. (2020) and Borsos et al. (2021) use infinite-width NTKs for BMDAL and related tasks. Han et al. (2021) propose sketching for infinite-width NTKs and also evaluate it for DAL. In contrast to these papers, we propose sketching for finite-width NTKs and allow combining the resulting kernel with different selection methods.

### 3.2 Selection Methods

Besides a Bayesian NN model, a Bayesian BMDAL method needs to specify an acquisition function that decides how to prioritize the pool samples. Many simple acquisition functions for quantifying uncertainty have been proposed (Kumar and Gupta, 2020). Selecting the next sample where an ensemble disagrees most is known as Query-by-Committee (QbC) (Seung et al., 1992). Krogh and Vedelsby (1994) employed QbC for DAL for regression. A more recent investigation of QbC to DAL for classification is performed by Beluch et al. (2018). Pop and Fulop (2018) combine ensembles with MC Dropout. Tsymbalov et al. (2018) use the predictive variance obtained by MC Dropout for DAL for regression. Zaverkin



and Kästner (2021) use last-layer-based uncertainty in DAL for regression on atomistic data. Unlike the other approaches mentioned before, the approach by Zaverkin and Kästner (2021) can be applied to a single NN trained without Dropout.

Many uncertainty-based acquisition functions do not distinguish between epistemic uncertainty, i.e., lack of knowledge about the true functional relationship, and aleatoric uncertainty, i.e., inherent uncertainty due to label noise. Hounsby et al. (2011) propose the BALD acquisition function, which aims to quantify epistemic uncertainty only. Gal et al. (2017) apply BALD and other acquisition functions to BMDAL for classification with MC Dropout. To enforce diversity of the selected batch, Kirsch et al. (2019) propose BatchBALD and evaluate it on classification problems with MC Dropout. Ash et al. (2021) propose BAIT, which also incorporates representativity through Fisher information based on last-layer features, and is evaluated on classification and regression data sets.

Another approach towards BMDAL is to find core-sets that represent  $\mathcal{X}_{\text{pool}}$  in a geometric sense. Sener and Savarese (2018) and Geifman and El-Yaniv (2017) propose algorithms to cover the pool set with  $\mathcal{X}_{\text{batch}} \cup \mathcal{X}_{\text{train}}$  in a last-layer feature space. Ash et al. (2019) propose BADGE, which applies clustering in a similar feature space, but includes uncertainty via gradients through the softmax layer for classification. ACS-FW (Pinsler et al., 2019) can be seen as a hybrid between core-set and Bayesian approaches, trying to approximate the expected log-posterior on the pool set with a core-set, also using last-layer-based Bayesian approximations. Besides BAIT, ACS-FW is one of the few approaches that is designed and evaluated for both classification and regression. Our newly proposed selection method LCMD is clustering-based like the k-means++ method used in BADGE, but deterministic.

Many more approaches towards BMDAL exist, and they can be combined with additional steps such as pre-reduction of  $\mathcal{X}_{\text{pool}}$  (Ghorbani et al., 2022) or re-weighting of selected instances (Farquhar et al., 2021). Most of these BMDAL methods are geared towards classification, and for a broader overview, we refer to Ren et al. (2021). For (image) regression, Ranganathan et al. (2020) introduce an auxiliary loss term on the pool set, which they use to perform DAL. It is unclear, though, to which extent their success is explained by implicitly performing semi-supervised learning.

Since we frequently consider Gaussian Processes (GPs) as approximations to Bayesian NNs in this paper, our work is also related to BMAL for GPs, although in our case the GPs are only used for selecting  $\mathcal{X}_{\text{batch}}$  and not for the regression itself. Popular BMAL methods for GPs have been suggested for example by Seo et al. (2000) and Krause et al. (2008).

### 3.3 Data Sets

In terms of benchmark data sets for BM(D)AL for regression, Tsymbalov et al. (2018) use seven large tabular data sets, some of which we have included in our benchmark, cf. Section E.1. Pinsler et al. (2019) use only one large tabular regression data set. Ash et al. (2021) use a small tabular regression data set and three image regression data sets, two of which are converted classification data sets. Wu (2018) benchmarks exclusively on small tabular data sets. Zaverkin and Kästner (2021) work with atomistic data sets, which require specialized NN architectures and longer training times, and are therefore less well-suited for a large-scale benchmark. Ranganathan et al. (2020) use CNNs on five image regression data sets. Recently, a benchmark for BMDAL for drug discovery has been proposed, which uses

four counterfactual regression data sets (Mehrjou et al., 2021). In this paper, we provide an open-source benchmark on 15 large tabular data sets, which includes more baselines and evaluation criteria than evaluations in previous papers.

## 4. Kernels

In this section, we discuss a variety of base kernels yielding various approximations to a trained NN  $f_{\theta_T}$ , as well as different kernel transformations that yield new kernels with different meanings or simply improved efficiency. In the following, we consider positive semi-definite kernels  $k : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ . For an introduction to kernels, we refer to the literature (e.g. Steinwart and Christmann, 2008). The kernels considered here can usually be represented by a feature map  $\phi$  with finite-dimensional feature space, that is,  $\phi : \mathbb{R}^d \rightarrow \mathbb{R}^{d_{\text{feat}}}$  with  $k(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$ . For a sequence  $\mathcal{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$  of inputs, which we sometimes treat like a set  $\mathcal{X} \subseteq \mathbb{R}^d$  by a slight abuse of notation, we define the corresponding feature matrix

$$\phi(\mathcal{X}) = \begin{pmatrix} \phi(\mathbf{x}_1)^\top \\ \vdots \\ \phi(\mathbf{x}_n)^\top \end{pmatrix} \in \mathbb{R}^{n \times d_{\text{feat}}} \quad (2)$$

and kernel matrices  $k(\mathbf{x}, \mathcal{X}) = (k(\mathbf{x}, \mathbf{x}_i))_i \in \mathbb{R}^{1 \times n}$ ,  $k(\mathcal{X}, \mathcal{X}) = (k(\mathbf{x}_i, \mathbf{x}_j))_{ij} \in \mathbb{R}^{n \times n}$ ,  $k(\mathcal{X}, \mathbf{x}) = (k(\mathbf{x}_i, \mathbf{x}))_i \in \mathbb{R}^{n \times 1}$ .

### 4.1 Base Kernels

We first discuss various options for creating base kernels that induce some notion of similarity on the training and pool inputs. An overview of these base kernels can be found in Table 1.

#### 4.1.1 LINEAR KERNEL

A very simple baseline for other base kernels is the linear kernel  $k_{\text{lin}}(\mathbf{x}, \tilde{\mathbf{x}}) = \langle \mathbf{x}, \tilde{\mathbf{x}} \rangle$ , corresponding to the identity feature map

$$\phi_{\text{lin}}(\mathbf{x}) := \mathbf{x} .$$

It is usually very fast to evaluate but does not represent the behavior of an NN well. Moreover, its feature space dimension depends on the input dimension, and hence may not be suited for selection methods that depend on having high-dimensional representations of the data. A more accurate representation of the behavior of an NN is given by the next kernel:

Base kernel	Symbol	Feature map	Feature space dimension $d_{\text{feat}}$
Linear	$k_{\text{lin}}$	$\phi_{\text{lin}}(\mathbf{x}) = \mathbf{x}$	$d$
NNGP	$k_{\text{nngp}}$	not explicitly defined	$\infty$
full gradient	$k_{\text{grad}}$	$\phi_{\text{grad}}(\mathbf{x}) = \nabla_{\theta} f_{\theta_T}(\mathbf{x})$	$\sum_{l=1}^L d_l(d_{l-1} + 1)$
last-layer	$k_{\text{ll}}$	$\phi_{\text{ll}}(\mathbf{x}) = \nabla_{\tilde{\mathbf{w}}^{(L)}} f_{\theta_T}(\mathbf{x})$	$d_L(d_{L-1} + 1)$

Table 1: Overview of the introduced base kernels.

## 4.1.2 FULL GRADIENT KERNEL

If  $\boldsymbol{\theta}_T$  is the parameter vector of the trained NN, we define

$$\phi_{\text{grad}}(\mathbf{x}) := \nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}_T}(\mathbf{x}) .$$

This is motivated as follows: A linearization of the NN with respect to its parameters around  $\boldsymbol{\theta}_T$  is given by the first-order Taylor expansion

$$f_{\boldsymbol{\theta}}(\mathbf{x}) \approx \tilde{f}_{\boldsymbol{\theta}}(\mathbf{x}) := f_{\boldsymbol{\theta}_T}(\mathbf{x}) + \langle \phi_{\text{grad}}(\mathbf{x}), \boldsymbol{\theta} - \boldsymbol{\theta}_T \rangle . \quad (3)$$

If we were to resume training from the parameters  $\boldsymbol{\theta}_T$  after labeling the next batch  $\mathcal{X}_{\text{batch}}$ , the result of training on the extended data could hence be approximated by the function  $f_{\boldsymbol{\theta}_T} + f_{\Delta}$ , where  $f_{\Delta}$  is the result of linear regression with feature map  $\phi_{\text{grad}}$  on the data residuals  $(\mathbf{x}_i, y_i - f_{\boldsymbol{\theta}_T}(\mathbf{x}_i))$  for  $(\mathbf{x}_i, y_i) \in \mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{batch}}$ .

The kernel  $k_{\text{grad}}$  is also known as the (empirical / finite-width) *neural tangent kernel* (NTK). It depends on the linearization point  $\boldsymbol{\theta}_T$ , but can for certain training settings converge to a fixed kernel as the hidden layer widths go to infinity (Jacot et al., 2018; Lee et al., 2019; Arora et al., 2019). In practical settings, however, it has been observed that  $k_{\text{grad}}$  often ‘‘improves’’ during training (Fort et al., 2020; Long, 2021; Shan and Bordelon, 2021; Atanasov et al., 2021), especially in the beginning of training. This agrees with our observations in Section 6 and suggests that shorter training might already yield a gradient kernel that allows selecting a good  $\mathcal{X}_{\text{batch}}$ . Indeed, Coleman et al. (2019) found that shorter training and even smaller models can already be sufficient to select good batches for BMDAL for classification.

For fully-connected layers, we will now show that the feature map  $\phi_{\text{grad}}$  has an additional product structure that can be exploited to reduce the runtime and memory consumption of a kernel evaluation. For notational simplicity, we rewrite Eq. (1) as

$$\begin{aligned} \mathbf{z}_i^{(l+1)} &= \tilde{\mathbf{W}}^{(l+1)} \tilde{\mathbf{x}}_i^{(l)}, \\ \tilde{\mathbf{W}}^{(l+1)} &:= (\mathbf{W}^{(l+1)} \quad \mathbf{b}^{(l+1)}) \in \mathbb{R}^{d_{l+1} \times (d_l+1)}, \quad \tilde{\mathbf{x}}_i^{(l)} := \begin{pmatrix} \frac{\sigma_w}{\sqrt{d_l}} \mathbf{x}_i^{(l)} \\ \sigma_b \end{pmatrix} \in \mathbb{R}^{d_l+1}, \end{aligned} \quad (4)$$

with parameters  $\boldsymbol{\theta} = (\tilde{\mathbf{W}}^{(1)}, \dots, \tilde{\mathbf{W}}^{(L)})$ . Using the notation from Eq. (4), we can write

$$\phi_{\text{grad}}(\mathbf{x}_i^{(0)}) = \left( \frac{d\mathbf{z}_i^{(L)}}{d\tilde{\mathbf{W}}^{(1)}}, \dots, \frac{d\mathbf{z}_i^{(L)}}{d\tilde{\mathbf{W}}^{(L)}} \right) = \left( \frac{d\mathbf{z}_i^{(L)}}{d\mathbf{z}_i^{(1)}} (\tilde{\mathbf{x}}_i^{(0)})^\top, \dots, \frac{d\mathbf{z}_i^{(L)}}{d\mathbf{z}_i^{(L)}} (\tilde{\mathbf{x}}_i^{(L-1)})^\top \right) . \quad (5)$$

For a kernel evaluation, the factorization of the weight matrix derivatives can be exploited via

$$\begin{aligned} k_{\text{grad}}(\mathbf{x}_i^{(0)}, \mathbf{x}_j^{(0)}) &= \sum_{l=1}^L \left\langle \frac{d\mathbf{z}_i^{(L)}}{d\mathbf{z}_i^{(l)}} (\tilde{\mathbf{x}}_i^{(l-1)})^\top, \frac{d\mathbf{z}_j^{(L)}}{d\mathbf{z}_j^{(l)}} (\tilde{\mathbf{x}}_j^{(l-1)})^\top \right\rangle_F \\ &= \sum_{l=1}^L \underbrace{\langle \tilde{\mathbf{x}}_i^{(l-1)}, \tilde{\mathbf{x}}_j^{(l-1)} \rangle}_{=: k_{\text{in}}^{(l)}(\mathbf{x}_i^{(0)}, \mathbf{x}_j^{(0)})} \cdot \underbrace{\left\langle \frac{d\mathbf{z}_i^{(L)}}{d\mathbf{z}_i^{(l)}}, \frac{d\mathbf{z}_j^{(L)}}{d\mathbf{z}_j^{(l)}} \right\rangle}_{=: k_{\text{out}}^{(l)}(\mathbf{x}_i^{(0)}, \mathbf{x}_j^{(0)})}, \end{aligned} \quad (6)$$

since  $\langle \mathbf{ab}^\top, \mathbf{cd}^\top \rangle_F = \text{tr}(\mathbf{ba}^\top \mathbf{cd}^\top) = \text{tr}(\mathbf{a}^\top \mathbf{cd}^\top \mathbf{b}) = \mathbf{a}^\top \mathbf{cd}^\top \mathbf{b} = \langle \mathbf{a}, \mathbf{c} \rangle \cdot \langle \mathbf{b}, \mathbf{d} \rangle$ . This means that  $k_{\text{grad}}$  can be decomposed into sums of products of kernels with smaller feature space dimension:<sup>2</sup>

$$k_{\text{grad}}(\mathbf{x}, \tilde{\mathbf{x}}) = \sum_{l=1}^L k_{\text{in}}^{(l)}(\mathbf{x}, \tilde{\mathbf{x}}) \cdot k_{\text{out}}^{(l)}(\mathbf{x}, \tilde{\mathbf{x}}) \quad (7)$$

When using Eq. (6), the full gradients  $\frac{dz_i^{(L)}}{d\tilde{\mathbf{w}}^{(l)}}$  never have to be computed or stored explicitly.

If  $\frac{dz_i^{(L)}}{dz_i^{(l)}}$  and  $\tilde{\mathbf{x}}_i^{(l-1)}$  are already computed and the hidden layers contain  $m = d_1 = \dots = d_{L-1}$  neurons each, Eq. (6) reduces the runtime complexity of a kernel evaluation from  $\Theta(m^2L)$  to  $\Theta(mL)$ , and similarly for the memory complexity of pre-computed features. In Section 4.2.3, we will see how to further accelerate this kernel computation using sketching. Efficient computations of  $k_{\text{grad}}$  for more general types of layers and multiple output neurons are discussed by Novak et al. (2022).

Since  $k_{\text{grad}}$  consists of gradient contributions from multiple layers, it is potentially important that the magnitudes of the gradients in different layers are balanced. We achieve this, at least at initialization, through the use of the neural tangent parameterization (Jacot et al., 2018). For other NN architectures, however, it might be desirable to re-weight gradient magnitudes from different layers to improve the results obtained with  $k_{\text{grad}}$ .

#### 4.1.3 LAST-LAYER KERNEL

A simple and rough approximation to the full-gradient kernel is given by only considering the gradient with respect to the parameters in the last layer:

$$\phi_{\Pi}(\mathbf{x}) := \nabla_{\tilde{\mathbf{w}}^{(L)}} f_{\theta_T}(\mathbf{x}) .$$

From Eq. (5), it is evident that in the single-output regression case that we are considering,  $\phi_{\Pi}(\mathbf{x}_i^{(0)})$  is simply the input  $\tilde{\mathbf{x}}_i^{(L-1)}$  to the last layer of the NN. The latter formulation can also be used in the multi-output setting, and versions of it (with  $\mathbf{x}_i^{(L-1)}$  instead of  $\tilde{\mathbf{x}}_i^{(L-1)}$ ) have been frequently used for BMDAL (Sener and Savarese, 2018; Geifman and El-Yaniv, 2017; Pinsler et al., 2019; Ash et al., 2019; Zaverkin and Kästner, 2021; Ash et al., 2021).

#### 4.1.4 INFINITE-WIDTH NNGP

It has been shown that as the widths  $d_1, \dots, d_{L-1}$  of the hidden NN layers converge to infinity, the distribution of the initial function  $f_{\theta_0}$  converges to a Gaussian Process with mean zero and a covariance kernel  $k_{\text{nngp}}$  called the neural network Gaussian process (NNGP) kernel (Neal, 1994; Lee et al., 2018; Matthews et al., 2018). This kernel depends on the network depth, the used activation function, and details such as the initialization variance and scaling factors like  $\sigma_w$ . In our experiments, we use the NNGP kernel corresponding to the employed NN setup, for which the formulas are given in Section B.1.

As mentioned above, there exists an infinite-width limit of  $k_{\text{grad}}$ , the so-called neural tangent kernel (Jacot et al., 2018). We decided to omit it from our experiments in Section E after preliminary experiments showed similarly bad performance as for the NNGP.

---

2. For the sketching method defined later, we may exploit that  $k_{\text{out}}^{(L)}(\mathbf{x}, \tilde{\mathbf{x}}) = 1$ , hence  $k_{\text{out}}^{(L)}$  can be omitted.

Notation	Description	$d_{\text{pre}}$	$d_{\text{post}}$	Configurable $\sigma^2$ ?
$k_{\rightarrow\text{scale}(\mathcal{X})}$	Rescale kernel to normalize mean $k(\mathbf{x}, \mathbf{x})$ on $\mathcal{X}$	any	$d_{\text{pre}}$	no
$k_{\rightarrow\text{post}(\mathcal{X}, \sigma^2)}$	GP posterior covariance after observing $\mathcal{X}$	any	$d_{\text{pre}}$	yes
$k_{\rightarrow\mathcal{X}}$	Short for $k_{\rightarrow\text{scale}(\mathcal{X})\rightarrow\text{post}(\mathcal{X}, \sigma^2)}$	any	$d_{\text{pre}}$	yes
$k_{\rightarrow\text{sketch}(p)}$	Sketching with $p$ features	$< \infty$	$p$	no
$k_{\rightarrow\text{ens}(N_{\text{ens}})}$	Sum of kernels for $N_{\text{ens}}$ ensembled networks	any	$N_{\text{ens}}d_{\text{pre}}$	no
$k_{\rightarrow\text{acs-grad}}$	Gradient-based kernel from Pinsler et al. (2019)	any	$d_{\text{pre}}^2$	yes
$k_{\rightarrow\text{acs-rf}(p)}$	Kernel from Pinsler et al. (2019) with $p$ random features	$< \infty$	$p$	yes
$k_{\rightarrow\text{acs-rf-hyper}(p)}$	Kernel from Pinsler et al. (2019) with $p$ random features and hyperprior on $\sigma^2$	$< \infty$	$p$	no

Table 2: Overview of our considered kernel transformations that can be applied to a kernel  $k$ . Here,  $d_{\text{pre}}$  refers to the feature space dimension of  $k$  and  $d_{\text{post}}$  refers to the feature space dimension after the transformation. Moreover,  $\sigma^2$  refers to the assumed noise variance in the GP model.

## 4.2 Kernel Transformations

The base kernels introduced in Section 4.1 are constructed such that kernel regression with these kernels serves as a proxy for regression with the corresponding NN. By using kernels, we can model interactions  $k(\mathbf{x}, \tilde{\mathbf{x}})$  between two inputs, which is crucial to incorporate diversity (DIV) into the selection methods. However, this is not always sufficient to apply a selection method. For example, sometimes we want the kernel to represent uncertainties of the NN after observing the data, or we want to reduce the feature space dimension to render selection more efficient. Therefore, we introduce various ways to transform kernels in this section. When applying transformations  $T_1, \dots, T_n$  in this order to a base kernel  $k_{\text{base}}$ , we denote the transformed kernel by  $k_{\text{base}\rightarrow T_1\rightarrow T_2\rightarrow \dots\rightarrow T_n}$ . Of course, we can only cover selected transformations relevant to our applications, and other transformations such as sums or products of kernels are possible as well.

### 4.2.1 SCALING

For a given kernel  $k$  with feature map  $\phi$  and scaling factor  $\lambda \in \mathbb{R}$ , we can construct the kernel  $\lambda^2 k$  with feature map  $\lambda\phi$ . This scaling can make a difference if we subsequently consider a Gaussian Process (GP) with covariance function  $\lambda^2 k$ . In this case,  $\lambda^2 k(\mathbf{x}, \tilde{\mathbf{x}})$  describes the covariance between  $f(\mathbf{x})$  and  $f(\tilde{\mathbf{x}})$  under the prior distribution over functions  $f$ . Since we train with normalized labels,  $N_{\text{train}}^{-1} \sum_{y \in \mathcal{Y}_{\text{train}}} y_i^2 \approx 1$ , we would like to choose the scaling factor  $\lambda$  such that  $N_{\text{train}}^{-1} \sum_{\mathbf{x} \in \mathcal{X}_{\text{train}}} \lambda^2 k(\mathbf{x}, \mathbf{x}) = 1$ . Therefore, we propose the automatic scale

normalization

$$k_{\rightarrow\text{scale}(\mathcal{X}_{\text{train}})}(\mathbf{x}, \tilde{\mathbf{x}}) := \lambda^2 k(\mathbf{x}, \tilde{\mathbf{x}}), \quad \lambda := \left( \frac{1}{N_{\text{train}}} \sum_{\mathbf{x} \in \mathcal{X}_{\text{train}}} k(\mathbf{x}, \mathbf{x}) \right)^{-1/2}.$$

#### 4.2.2 GAUSSIAN PROCESS POSTERIOR TRANSFORMATION

For a given kernel  $k$  with corresponding feature map  $\phi$ , we can consider a Gaussian Process (GP) with kernel  $k$ , which is equivalent to a Bayesian linear regression model with feature map  $\phi$ : In feature space, we model our observations as  $y_i = \mathbf{w}^\top \phi(\mathbf{x}_i) + \varepsilon_i$  with weight prior  $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$  and i.i.d. observation noise  $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ . The random function  $f(\mathbf{x}_i) := \mathbf{w}^\top \phi(\mathbf{x}_i)$  now has the covariance function  $\text{Cov}(f(\mathbf{x}_i), f(\mathbf{x}_j)) = \phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_j) = k(\mathbf{x}_i, \mathbf{x}_j)$ .

It is well-known, see e.g. Section 2.1 and 2.2 in Bishop (2006), that the posterior distribution of a Gaussian process after observing the training data  $\mathcal{D}_{\text{train}}$  with inputs  $\mathcal{X}_{\text{train}}$  is also a Gaussian process with kernel

$$\begin{aligned} k_{\rightarrow\text{post}(\mathcal{X}_{\text{train}}, \sigma^2)}(\mathbf{x}, \tilde{\mathbf{x}}) &:= \text{Cov}(f(\mathbf{x}), f(\tilde{\mathbf{x}}) \mid \mathcal{X}_{\text{train}}, \mathcal{Y}_{\text{train}}) \\ &= k(\mathbf{x}, \tilde{\mathbf{x}}) - k(\mathbf{x}, \mathcal{X}_{\text{train}})(k(\mathcal{X}_{\text{train}}, \mathcal{X}_{\text{train}}) + \sigma^2 \mathbf{I})^{-1} k(\mathcal{X}_{\text{train}}, \tilde{\mathbf{x}}) \quad (8) \\ &\stackrel{\text{see below}}{=} \phi(\mathbf{x})^\top (\sigma^{-2} \phi(\mathcal{X}_{\text{train}})^\top \phi(\mathcal{X}_{\text{train}}) + \mathbf{I})^{-1} \phi(\tilde{\mathbf{x}}) \quad (9) \\ &= \sigma^2 \phi(\mathbf{x})^\top (\phi(\mathcal{X}_{\text{train}})^\top \phi(\mathcal{X}_{\text{train}}) + \sigma^2 \mathbf{I})^{-1} \phi(\tilde{\mathbf{x}}). \quad (10) \end{aligned}$$

Here, the equivalence between Eq. (8) and Eq. (9) for  $\sigma^2 > 0$  can be obtained using the Woodbury matrix identity. In our implementation, we use the feature map version, Eq. (9), whenever  $d_{\text{feat}} \leq \max\{1024, 3|\mathcal{X}_{\text{train}}|\}$ . An explicit feature map can be obtained from Eq. (10) as

$$\phi_{\rightarrow\text{post}(\mathcal{X}_{\text{train}}, \sigma^2)}(\mathbf{x}) = \sigma (\phi(\mathcal{X}_{\text{train}})^\top \phi(\mathcal{X}_{\text{train}}) + \sigma^2 \mathbf{I})^{-1/2} \phi(\mathbf{x}).$$

If the posterior with respect to two disjoint sets of inputs  $\mathcal{X}_1, \mathcal{X}_2 \subseteq \mathbb{R}^d$  is sought, it is equivalent to condition first on  $\mathcal{X}_1$  and then on  $\mathcal{X}_2$ :

$$k_{\rightarrow\text{post}(\mathcal{X}_1 \cup \mathcal{X}_2, \sigma^2)}(\mathbf{x}, \tilde{\mathbf{x}}) = k_{\rightarrow\text{post}(\mathcal{X}_1, \sigma^2) \rightarrow \text{post}(\mathcal{X}_2, \sigma^2)}(\mathbf{x}, \tilde{\mathbf{x}}). \quad (11)$$

In our experiments, we rescale kernels before applying the posterior transformation, which we abbreviate by

$$k_{\rightarrow\mathcal{X}_{\text{train}}}(\mathbf{x}, \tilde{\mathbf{x}}) := k_{\rightarrow\text{scale}(\mathcal{X}_{\text{train}}) \rightarrow \text{post}(\mathcal{X}_{\text{train}}, \sigma^2)}(\mathbf{x}, \tilde{\mathbf{x}}).$$

The application of the posterior transformation to network-dependent kernels can be seen as an instance of approximate inference with Bayesian NNs. Specifically, we show in Section C.1 that  $k_{\text{ll} \rightarrow \text{post}(\mathcal{X}_{\text{train}}, \sigma^2)}$  and  $k_{\text{grad} \rightarrow \text{post}(\mathcal{X}_{\text{train}}, \sigma^2)}$  correspond to last-layer and generalized Gauss-Newton (GGN) approximations to the Hessian in a Laplace approximation (Laplace, 1774; MacKay, 1992a) for Bayesian NNs, see also Khan et al. (2019). Moreover, in the case of the last-layer kernel, this procedure is equivalent to interpreting the last layer of the NN as a Bayesian linear regression model.

### 4.2.3 SKETCHING

Sketching methods, which allow approximating matrices like  $\phi(\mathcal{X})$  with smaller matrices in some sense, can be used to approximate a kernel  $k$  with high-dimensional feature space by a kernel with a lower-dimensional feature space (see e.g. Woodruff, 2014). For example,  $k_{\text{grad}}$  and kernels resulting from the ACS gradient transformation introduced in Section 4.2.6 involve product kernels with very high-dimensional feature spaces ( $d_{\text{feat}} > 250,000$ ). In our experiments, we apply sketching mainly to these kernels. This is especially useful for methods such as the posterior transformation discussed previously and the FRANKWOLFE and BAIT selection methods explained in Section 5.2, which are not very efficient in the kernel formulation.

We sketch finite-dimensional feature maps as follows:

- (1) **Generic finite-dimensional feature maps:** Consider a generic kernel  $k$  with finite-dimensional feature map  $\phi : \mathbb{R}^{d_0} \rightarrow \mathbb{R}^{d_{\text{pre}}}$ . For a random vector  $\mathbf{u} \sim \mathcal{N}(0, \mathbf{I}_{d_{\text{pre}}})$ , a single random feature is given by the feature map  $\phi_{\mathbf{u}}(\mathbf{x}) := \mathbf{u}^\top \phi(\mathbf{x})$ , which yields an unbiased estimate of the kernel since  $\mathbb{E}_{\mathbf{u}} \langle \phi_{\mathbf{u}}(\mathbf{x}), \phi_{\mathbf{u}}(\tilde{\mathbf{x}}) \rangle = k(\mathbf{x}, \tilde{\mathbf{x}})$ . By combining multiple such random features, the accuracy of the kernel approximation can be improved. For  $p$  random features, we obtain the random feature map

$$\phi_{\rightarrow\text{sketch}(p)}(\mathbf{x}) := \frac{1}{\sqrt{p}} \mathbf{U} \phi(\mathbf{x}) \in \mathbb{R}^p, \quad (12)$$

where  $\mathbf{U} \in \mathbb{R}^{p \times d_{\text{pre}}}$  is a random matrix with i.i.d. standard normal entries. This is also known as a Gaussian sketch.

In terms of the kernel distance

$$d_k(\mathbf{x}, \tilde{\mathbf{x}}) := \|\phi(\mathbf{x}) - \phi(\tilde{\mathbf{x}})\|_2 = \sqrt{k(\mathbf{x}, \mathbf{x}) + k(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}) - 2k(\mathbf{x}, \tilde{\mathbf{x}})}, \quad (13)$$

the approximation quality of the sketched kernel can be analyzed using variants of the celebrated Johnson-Lindenstrauss lemma (Johnson and Lindenstrauss, 1984). For example, the following variant is proved in Section C.2 based on a result by Arriaga and Vempala (1999).

**Theorem 1** (Variant of the Johnson-Lindenstrauss Lemma). *Let  $\varepsilon, \delta \in (0, 1)$  and let  $\mathcal{X} \subseteq \mathbb{R}^d$  be finite. If*

$$p \geq 8 \log(|\mathcal{X}|^2 / \delta) / \varepsilon^2, \quad (14)$$

*then the following bound on all pairwise distances holds with probability  $\geq 1 - \delta$  for the Gaussian sketch in Eq. (12):*

$$\forall \mathbf{x}, \tilde{\mathbf{x}} \in \mathcal{X} : (1 - \varepsilon) d_k(\mathbf{x}, \tilde{\mathbf{x}}) \leq d_{k_{\rightarrow\text{sketch}(p)}}(\mathbf{x}, \tilde{\mathbf{x}}) \leq (1 + \varepsilon) d_k(\mathbf{x}, \tilde{\mathbf{x}}). \quad (15)$$

Note that, counterintuitively, the lower bound in Eq. (14) does not depend on the feature space dimension  $d_{\text{pre}}$  of  $k$ .

- (2) **Sums of kernels:** Consider the sum kernel  $k = k_1 + k_2$  of kernels  $k_1, k_2$  with finite-dimensional feature maps  $\phi_1, \phi_2$ . A feature map for  $k$  is given by  $\phi(\mathbf{x}) := \begin{pmatrix} \phi_1(\mathbf{x}) \\ \phi_2(\mathbf{x}) \end{pmatrix}$ .

We can apply sketching as

$$\phi_{\rightarrow\text{sketch}(p)}(\mathbf{x}) := \phi_{1 \rightarrow \text{sketch}(p)}(\mathbf{x}) + \phi_{2 \rightarrow \text{sketch}(p)}(\mathbf{x}). \quad (16)$$

This again yields an unbiased estimate of the kernel  $k$ . If  $\phi_{1 \rightarrow \text{sketch}(p)}$  and  $\phi_{2 \rightarrow \text{sketch}(p)}$  are sketched as in Eq. (12), then Eq. (16) is equivalent to sketching  $\phi$  with Eq. (12) directly.

- (3) **Products of kernels:** Consider the product kernel  $k = k_1 \cdot k_2$  of kernels  $k_1, k_2$  with finite-dimensional feature maps  $\phi_1, \phi_2$ . A feature map for  $k$  is given by  $\phi(\mathbf{x}) := \phi_1(\mathbf{x}) \otimes \phi_2(\mathbf{x})$ , where  $\otimes$  is the tensor product. Hence, if the feature spaces of  $\phi_1$  and  $\phi_2$  have dimensions  $p_1$  and  $p_2$ , respectively, the feature space of  $\phi$  has dimension  $p_1 p_2$ . While this dimension is still finite, using Eq. (12) for sketching would potentially require a large amount of memory for storing  $\mathbf{U}$  as well as a large runtime for the matrix-vector product. Therefore, we sketch product kernels more efficiently as

$$\phi_{\rightarrow \text{sketch}(p)}(\mathbf{x}) := \sqrt{p} \phi_{1 \rightarrow \text{sketch}(p)}(\mathbf{x}) \odot \phi_{2 \rightarrow \text{sketch}(p)}(\mathbf{x}), \quad (17)$$

where  $\odot$  denotes the element-wise product (or Hadamard product). This again yields an unbiased estimator of  $k$  without the need to perform computations in the  $p_1 p_2$ -dimensional feature space. While this simple tensor sketching method works sufficiently well for our purposes, its approximation properties are suboptimal and can be improved with a more complicated sketching method (Ahle et al., 2020).

In the kernel  $k_{\text{grad} \rightarrow \text{sketch}(p) \rightarrow \text{post}(\mathcal{X}_{\text{train}}, \sigma^2)}$ , the inclusion of sketching can be considered a further approximation to the posterior predictive distribution for Bayesian NNs. In this context, a different sketching method has been proposed by Sharma et al. (2021). It is also possible to apply sketching to kernels with infinite-dimensional feature space (see e.g. Kar and Karnick, 2012; Zandieh et al., 2021; Han et al., 2021), but such kernels are less relevant in our case.

#### 4.2.4 ENSEMBLING

Ensembles of NNs have been demonstrated to yield good uncertainty estimates for DAL (Beluch et al., 2018) and can improve the uncertainty estimates of MC Dropout (Pop and Fulop, 2018). This motivates the study of ensembled kernels. When multiple NNs are trained on the same data and a kernel  $k^{(i)}$  is computed for each model  $i \in \{1, \dots, N_{\text{ens}}\}$  via a base kernel and a list of transformations, these kernels can be ensembled simply by adding them together:

$$k_{\rightarrow \text{ens}(N_{\text{ens}})} = k^{(1)} + \dots + k^{(N_{\text{ens}})}.$$

In the context of Bayesian NNs, ensembling of posterior kernels is related to a mixture of Laplace approximations (Eschenhagen et al., 2021), cf. Section C.

#### 4.2.5 ACS RANDOM FEATURES TRANSFORMATION

In the following two paragraphs, we will briefly introduce multiple kernel transformations corresponding to various alternative ways of applying the ACS-FW method by Pinsler et al. (2019) to GP regression. For a more complete description, we refer to the original publication. ACS-FW seeks to approximate the expected complete data log posterior,  $f_{\text{pool}}(\boldsymbol{\theta}) = \mathbb{E}_{\mathcal{Y}_{\text{pool}} \sim P(\mathcal{Y}_{\text{pool}} | \mathcal{X}_{\text{pool}}, \mathcal{D}_{\text{train}})} \log p(\boldsymbol{\theta} | \mathcal{D}_{\text{train}}, \mathcal{X}_{\text{pool}}, \mathcal{Y}_{\text{pool}})$ , with the expected log posterior of the train data and the next batch,  $f_{\text{batch}}(\boldsymbol{\theta}) = \mathbb{E}_{\mathcal{Y}_{\text{batch}} \sim P(\mathcal{Y}_{\text{batch}} | \mathcal{X}_{\text{batch}}, \mathcal{D}_{\text{train}})} \log p(\boldsymbol{\theta} |$



$\mathcal{D}_{\text{train}}, \mathcal{X}_{\text{batch}}, \mathcal{Y}_{\text{batch}}$ ). Here, the labels  $\mathcal{Y}_{\text{pool}}$  and  $\mathcal{Y}_{\text{batch}}$  are drawn from the posterior distribution after observing  $\mathcal{D}_{\text{train}}$ . For a given Bayesian model, they then define different kernels resulting from this objective. As a Bayesian model, we use the same Gaussian process model as for the posterior transformation above, with kernel  $k_{\rightarrow\text{scale}}(\mathcal{X}_{\text{train}})$ . In this case, as shown in Section C.3, we have  $f_{\text{pool}}(\boldsymbol{\theta}) - f_{\text{batch}}(\boldsymbol{\theta}) = \sum_{\mathbf{x} \in \mathcal{X}_{\text{pool}} \setminus \mathcal{X}_{\text{batch}}} f_{\text{acs}}(\mathbf{x}, \boldsymbol{\theta})$  with

$$f_{\text{acs}}(\mathbf{x}, \boldsymbol{\theta}) := \frac{1}{2} \log \left( 1 + \frac{k_{\rightarrow\mathcal{X}_{\text{train}}}(\mathbf{x}, \mathbf{x})}{\sigma^2} \right) - \frac{(\boldsymbol{\theta}^\top \phi_{\rightarrow\text{scale}}(\mathcal{X}_{\text{train}})(\mathbf{x}))^2 + k_{\rightarrow\mathcal{X}_{\text{train}}}(\mathbf{x}, \mathbf{x})}{2\sigma^2}. \quad (18)$$

The weighted inner product by (Pinsler et al., 2019) can then be written as

$$k_{\rightarrow\text{acs}}(\mathbf{x}, \tilde{\mathbf{x}}) := \mathbb{E}_{\boldsymbol{\theta} \sim P(\boldsymbol{\theta} | \mathcal{D}_{\text{train}})} [f_{\text{acs}}(\mathbf{x}, \boldsymbol{\theta}) f_{\text{acs}}(\tilde{\mathbf{x}}, \boldsymbol{\theta})]. \quad (19)$$

The expectation in Eq. (19) can be approximated using Monte Carlo quadrature as

$$k_{\rightarrow\text{acs}}(\mathbf{x}, \tilde{\mathbf{x}}) \approx k_{\rightarrow\text{acs-rf}}(d_{\text{post}}) := \frac{1}{d_{\text{post}}} \sum_{i=1}^{d_{\text{post}}} [f_{\text{acs}}(\mathbf{x}, \boldsymbol{\theta}^{(i)}) f_{\text{acs}}(\tilde{\mathbf{x}}, \boldsymbol{\theta}^{(i)})],$$

where  $\boldsymbol{\theta}^{(i)} \sim P(\boldsymbol{\theta} | \mathcal{D}_{\text{train}})$  are i.i.d. parameter samples from the posterior. This corresponds to the random features approximation proposed by Pinsler et al. (2019), given by

$$\phi_{\rightarrow\text{acs-rf}}(d_{\text{post}})(\mathbf{x}) := d_{\text{post}}^{-1/2} (f_{\text{acs}}(\mathbf{x}, \boldsymbol{\theta}^{(1)}), \dots, f_{\text{acs}}(\mathbf{x}, \boldsymbol{\theta}^{(d_{\text{post}})}))^\top.$$

In their regression experiments, Pinsler et al. (2019) use a slightly different feature map which we denote by  $\phi_{\rightarrow\text{acs-rf-hyper}}(d_{\text{post}})$  in our experiments. They state that this has been derived from a GP model with a hyperprior on  $\sigma^2$ , although no hints on its derivation are provided in their paper, so we directly use their source code for our implementation.

#### 4.2.6 ACS GRADIENT TRANSFORMATION

As an alternative to the random features approximation in the previous paragraph, Pinsler et al. (2019) proposed the *weighted Fisher inner product* given by

$$k_{\rightarrow\text{acs-grad}}(\mathbf{x}, \tilde{\mathbf{x}}) := \mathbb{E}_{\boldsymbol{\theta} \sim P(\boldsymbol{\theta} | \mathcal{D}_{\text{train}})} [\langle \nabla_{\boldsymbol{\theta}} f_{\text{acs}}(\mathbf{x}, \boldsymbol{\theta}), \nabla_{\boldsymbol{\theta}} f_{\text{acs}}(\tilde{\mathbf{x}}, \boldsymbol{\theta}) \rangle]. \quad (20)$$

Under the Gaussian process model, they showed that an explicit formula for  $k_{\rightarrow\text{acs-grad}}$  is given by

$$k_{\rightarrow\text{acs-grad}}(\mathbf{x}, \tilde{\mathbf{x}}) = \frac{1}{\sigma^4} k_{\rightarrow\text{scale}}(\mathcal{X}_{\text{train}})(\mathbf{x}, \tilde{\mathbf{x}}) k_{\rightarrow\mathcal{X}_{\text{train}}}(\mathbf{x}, \tilde{\mathbf{x}}).$$

Since this product kernel can have a high-dimensional feature space, they used  $k_{\rightarrow\text{acs-grad}}$  only on small data sets. In our experiments, we apply sketching to this kernel to scale it to large data sets. Again, they specified that they included a hyperprior on  $\sigma^2$  in their experiments, but their corresponding implementation appears to be equivalent to  $k_{\rightarrow\text{acs-grad}}$  for our purposes.

### 4.3 Discussion

Out of the base kernels and kernel transformations considered above, the training labels  $\mathcal{Y}_{\text{train}}$  only influence the base kernels  $k_{\parallel}$  and  $k_{\text{grad}}$  through the trained parameters  $\boldsymbol{\theta}_T$ , and to some extent the transformation  $k_{\rightarrow\text{acs-rf-hyper}(p)}$ . Using  $k_{\text{lin}}$  and  $k_{\text{nngp}}$  with the selection methods below thus leads to passive learning or experimental design, where the entire set of training inputs  $\mathcal{X}_{\text{train}}$  is selected before any of the labels  $\mathcal{Y}_{\text{train}}$  are computed. This can be much cheaper because no NN retraining is required, but is also potentially less accurate.

Another consideration to be made when selecting a kernel is the feature space dimension  $d_{\text{feat}}$ . While a low  $d_{\text{feat}}$  is usually beneficial for runtime purposes, larger  $d_{\text{feat}}$  might allow for a more accurate representation of over-parameterized NNs. For  $k_{\text{lin}}$ ,  $d_{\text{feat}}$  depends on the data set but is often rather small. For  $k_{\parallel}$ ,  $d_{\text{feat}}$  can be reduced using sketching, but an effective increase of  $d_{\text{feat}}$  requires increasing the width of the last hidden layer, which might not always be desirable. For  $k_{\text{grad}}$ ,  $d_{\text{feat}}$  is typically very large, and can be flexibly adjusted by using sketching. For  $k_{\text{nngp}}$ , we have  $d_{\text{feat}} = \infty$ , but sketching may also be applicable, see Han et al. (2021) for a similar application to infinite-width NTKs.

## 5. Selection Methods

In the following, we will discuss a variety of kernel-based selection methods. We first introduce the general iterative scheme that all evaluated methods (except BAIT-FB) use, with its two variants called P (for pool) and TP (for train+pool). Subsequently, we explain specific selection methods.

### 5.1 Iterative Selection Methods

A natural approach towards selecting  $\mathcal{X}_{\text{batch}}$  is to formulate an acquisition function  $a$  which scores an entire batch, such that  $a(\mathcal{X}_{\text{batch}})$  should be maximized over all  $\mathcal{X}_{\text{batch}} \subseteq \mathcal{X}_{\text{pool}}$  of size  $N_{\text{batch}}$ . However, the corresponding optimization problem is often intractable (Gonzalez, 1985; Civril and Magdon-Ismail, 2013). Many BMAL methods thus select points in a greedy/iterative fashion. To favor samples with high informativeness in an iterative selection scheme that tries to enforce diversity of the selected batch, two approaches can be used:

- (P) Informativeness can be incorporated through the kernel. For example,  $k_{\rightarrow\mathcal{X}_{\text{train}}}(\mathbf{x}, \mathbf{x})$  represents the posterior variance at  $\mathbf{x}$  of a GP with scaled kernel  $k_{\rightarrow\text{scale}(\mathcal{X}_{\text{train}})}$ .
- (TP) Informativeness can be incorporated implicitly by enforcing diversity of  $\mathcal{X}_{\text{train}} \cup \mathcal{X}_{\text{batch}}$  instead of only enforcing diversity of  $\mathcal{X}_{\text{batch}}$ . In other words, a batch that is sufficiently different from the training set typically necessarily contains new information.

An iterative selection template with the two variants P and TP is shown in Algorithm 3, where different choices of NEXTBATCH lead to different selection methods as discussed in Section 5.2. For simplicity of notation, Algorithm 3 does not reuse information in subsequent calls to NEXTBATCH, which however is necessary to make the selection methods more efficient. We provide efficiency-focused pseudocode, which is also used for our implementation, and an analysis of runtime and memory complexities in Section D. Additionally, our implementation usually accelerates kernel computations through suitable precomputations, often by precomputing the features  $\phi(\mathbf{x})$  for  $\mathbf{x} \in \mathcal{X}_{\text{train}} \cup \mathcal{X}_{\text{pool}}$ . In our notation, we treat  $\mathcal{X}_{\text{train}}$

---

**Algorithm 3** Iterative selection algorithm template with customizable function NEXTSAMPLE, for which different options will be discussed in Section 5.2.

---

```

function SELECT( $k$ ,  $\mathcal{X}_{\text{train}}$ ,  $\mathcal{X}_{\text{pool}}$ ,  $N_{\text{batch}}$ ,  $\text{mode} \in \{\text{P}, \text{TP}\}$ )
     $\mathcal{X}_{\text{batch}} \leftarrow \emptyset$ 
     $\mathcal{X}_{\text{mode}} \leftarrow \mathcal{X}_{\text{train}}$  if  $\text{mode} = \text{TP}$  else  $\emptyset$  ▷ Points considered as “selected”
    for  $i$  from 1 to  $N_{\text{batch}}$  do
         $\mathcal{X}_{\text{sel}} \leftarrow \mathcal{X}_{\text{mode}} \cup \mathcal{X}_{\text{batch}}$  ▷ Currently “selected” points
         $\mathcal{X}_{\text{rem}} \leftarrow \mathcal{X}_{\text{pool}} \setminus \mathcal{X}_{\text{batch}}$  ▷ Currently unselected points
         $\mathcal{X}_{\text{batch}} \leftarrow \mathcal{X}_{\text{batch}} \cup \{\text{NEXTSAMPLE}(k, \mathcal{X}_{\text{sel}}, \mathcal{X}_{\text{rem}})\}$ 
    end for
    return  $\mathcal{X}_{\text{batch}}$ 
end function

```

---

and  $\mathcal{X}_{\text{pool}}$  as sets, assuming that all values are distinct. In practice, if multiple identical  $\mathbf{x}_i$  are contained in  $\mathcal{X}_{\text{train}}$  and/or  $\mathcal{X}_{\text{pool}}$ , they should still be treated as distinct.

## 5.2 Specific Methods

In the following, we will discuss a variety of choices for NEXTSAMPLE in Algorithm 3, leading to different selection methods. An overview of the resulting selection methods is given in Table 3. Table 4 shows how BMAL methods from the literature relate to the presented selection methods and kernels.

### 5.2.1 RANDOM SELECTION

A simple baseline for comparison to other selection methods, denoted as RANDOM, is to select  $\mathcal{X}_{\text{batch}}$  randomly. We can formally express this as

$$\text{NEXTSAMPLE}(k, \mathcal{X}_{\text{sel}}, \mathcal{X}_{\text{rem}}) \sim \mathcal{U}(\mathcal{X}_{\text{rem}}) ,$$

where  $\mathcal{U}(\mathcal{X}_{\text{rem}})$  is the uniform distribution over  $\mathcal{X}_{\text{rem}}$ . Since NEXTSAMPLE does not use  $\mathcal{X}_{\text{sel}}$ , the P and TP versions of RANDOM are equivalent.

### 5.2.2 NAIVE ACTIVE LEARNING

If  $k(\mathbf{x}, \mathbf{x})$  is interpreted as a measure for the uncertainty of the model at  $\mathbf{x}$ , naive active learning can simply be formalized as

$$\text{NEXTSAMPLE}(k, \mathcal{X}_{\text{sel}}, \mathcal{X}_{\text{rem}}) = \underset{\mathbf{x} \in \mathcal{X}_{\text{rem}}}{\text{argmax}} k(\mathbf{x}, \mathbf{x}) . \tag{21}$$

Since NEXTSAMPLE only considers the diagonal of the kernel matrix  $k(\mathcal{X}_{\text{rem}}, \mathcal{X}_{\text{rem}})$ , we call the corresponding selection method MAXDIAG. Similar to RANDOM, the P and TP versions of MAXDIAG are equivalent. If  $k = \tilde{k}_{\rightarrow \mathcal{X}_{\text{train}}}$ ,  $k(\mathbf{x}, \mathbf{x}) + \sigma^2$  represents the posterior predictive variance of a GP with kernel  $\tilde{k}_{\rightarrow \text{scale}(\mathcal{X}_{\text{train}})}$  at  $\mathbf{x}$ . Unlike in the classification setting, the noise distribution  $\varepsilon \sim \mathcal{N}(0, \sigma^2)$  in the GP model is independent of  $\mathbf{x}$ , which renders different acquisition functions like maximum entropy (Shannon, 1948; MacKay, 1992b) or BALD (Houlsby et al., 2011) equivalent to Eq. (21). The active learning approach proposed by

Selection method	Description	Runtime complexity
RANDOM	Random selection	$\mathcal{O}(N_{\text{pool}} \log N_{\text{pool}})$
MAXDIAG	Naive active learning, picking largest diagonal entries	$\mathcal{O}(N_{\text{pool}}(T_k + \log N_{\text{pool}}))$
MAXDET	Greedy determinant maximization	$\mathcal{O}(N_{\text{cand}}N_{\text{sel}}(T_k + N_{\text{sel}}))$ or $\mathcal{O}(N_{\text{cand}}N_{\text{sel}}d_{\text{feat}})$
BAIT-F	Forward-Greedy total uncertainty minimization	$\mathcal{O}(N_{\text{cand}}N_{\text{sel}}d_{\text{feat}} + (N_{\text{train}} + N_{\text{pool}})d_{\text{feat}}^2)$
BAIT-FB	Forward-Backward-Greedy total uncertainty minimization	$\mathcal{O}(N_{\text{cand}}N_{\text{sel}}d_{\text{feat}} + (N_{\text{train}} + N_{\text{pool}})d_{\text{feat}}^2)$
FRANKWOLFE	Approximate kernel mean embedding using Frank-Wolfe	$\mathcal{O}((N_{\text{cand}} + N_{\text{pool}}N_{\text{batch}})d_{\text{feat}})$ or $\mathcal{O}(N_{\text{cand}}^2(T_k + 1))$
MAXDIST	Greedy distance maximization	$\mathcal{O}(N_{\text{pool}}N_{\text{sel}}(T_k + 1))$
KMEANSPP	Next point probability proportional to squared distance	$\mathcal{O}(N_{\text{pool}}N_{\text{sel}}(T_k + 1))$
LCMD (ours)	Greedy distance maximization in largest cluster	$\mathcal{O}(N_{\text{pool}}N_{\text{sel}}(T_k + 1))$

Table 3: Selection methods presented in this paper and their runtime complexities. For the runtime notation, we let  $T_k$  denote the runtime of a kernel evaluation and  $d_{\text{feat}}$  the dimensionality of its (pre-computed) features. Moreover, we write  $N_{\text{cand}} := N_{\text{pool}} + |\mathcal{X}_{\text{mode}}|$  and  $N_{\text{sel}} := N_{\text{batch}} + |\mathcal{X}_{\text{mode}}|$ , with  $\mathcal{X}_{\text{mode}}$  as in Algorithm 3. The runtime complexities are derived in Section D. Further refinements of the runtime complexities for RANDOM and MAXDIAG are possible but not practically relevant to us, as these methods are already very efficient.

Zaverkin and Kästner (2021) corresponds to applying MAXDIAG to  $k_{\parallel \rightarrow \mathcal{X}_{\text{train}}}$  in the limit  $\sigma^2 \rightarrow 0$ . Out of the three objectives presented in Section 2.2, MAXDIAG satisfies (INF), but not (DIV) and (REP). Indeed, if the pool set contains (almost-) duplicates, MAXDIAG may select a batch consisting of (almost) identical inputs.

### 5.2.3 GREEDY DETERMINANT MAXIMIZATION

To take account of the inputs  $\mathcal{X}_{\text{sel}}$  that have already been selected, it is possible to also condition the GP on the selected values  $\mathcal{X}_{\text{sel}}$ , since the posterior variance of a GP does not depend on the unknown labels for  $\mathcal{X}_{\text{sel}}$ . Picking the input with maximal uncertainty after conditioning is equivalent to maximizing a determinant, as we show in Section D.4:

$$\begin{aligned}
 \text{NEXTSAMPLE}(k, \mathcal{X}_{\text{sel}}, \mathcal{X}_{\text{rem}}) &= \operatorname{argmax}_{\mathbf{x} \in \mathcal{X}_{\text{rem}}} k_{\rightarrow \text{post}(\mathcal{X}_{\text{sel}}, \sigma^2)}(\mathbf{x}, \mathbf{x}) \\
 &= \operatorname{argmax}_{\mathbf{x} \in \mathcal{X}_{\text{rem}}} \det(k(\mathcal{X}_{\text{sel}} \cup \{\mathbf{x}\}, \mathcal{X}_{\text{sel}} \cup \{\mathbf{x}\}) + \sigma^2 \mathbf{I}), \quad (22)
 \end{aligned}$$

Known as	Selection method	Kernel	Remark
BALD (Houlsby et al., 2011)	MAXDIAG	$k_{\rightarrow\text{post}}(\mathcal{X}_{\text{train}}, \sigma^2)$	for GP with kernel $k$
BatchBALD (Kirsch et al., 2019)	MAXDET-P	$k_{\rightarrow\text{post}}(\mathcal{X}_{\text{train}}, \sigma^2)$	for GP with kernel $k$ , proposed for classification
BAIT (Ash et al., 2021)	BAIT-FB-P	$k_{\parallel\rightarrow\text{post}}(\mathcal{X}_{\text{train}}, \sigma^2)$	
ACS-FW (Pinsler et al., 2019)	FRANKWOLFE-P	$k_{\parallel\rightarrow\text{acs-rf}(p)}$ or $k_{\parallel\rightarrow\text{acs-grad}}$ or $k_{\parallel\rightarrow\text{acs-rf-hyper}(p)}$	
Core-Set (Sener and Savarese, 2018)	MAXDIST-TP*	similar to $k_{\parallel}$	proposed for classification
FF-Active (Geifman and El-Yaniv, 2017)	MAXDIST-TP	similar to $k_{\parallel}$	proposed for classification
BADGE (Ash et al., 2019)	KMEANSPP-P	similar to $k_{\parallel}$	proposed for classification

\* This refers to their simpler k-center-greedy selection method.

Table 4: Some (regression adaptations of) BM(D)AL methods from the literature and their corresponding selection methods and kernels.

We call the corresponding selection method MAXDET. It is equivalent to performing non-batch mode active learning on the GP with kernel  $k$ , and has been applied to GPs by Seo et al. (2000). Moreover, as we show in Section D.4, it is also equivalent to applying BatchBALD (Kirsch et al., 2019) to the GP with kernel  $k$ . If  $\sigma^2 = 0$ , it is equivalent to the  $P$ -greedy method for kernel interpolation (De Marchi et al., 2005), and it is also related to the greedy algorithm for D-optimal design (Wynn, 1970). In comparison to a naive implementation that computes each determinant separately, the runtime complexity of the determinant computation in MAXDET can be reduced by a factor of  $\mathcal{O}(N_{\text{sel}}^2)$  to  $\mathcal{O}(N_{\text{cand}}N_{\text{sel}}(T_k + N_{\text{sel}}))$  when implementing MAXDET via a partial pivoted matrix-free Cholesky decomposition as suggested in Pazouki and Schaback (2011). For the case  $N_{\text{sel}} \gg d_{\text{feat}}$ , we show in Section D.4 how the runtime complexity can be reduced further. For given  $\sigma^2 > 0$  in Eq. (22), it follows from Eq. (11) that applying MAXDET-TP to a kernel  $k$  is equivalent to applying MAXDET-P to  $k_{\rightarrow\text{post}}(\mathcal{X}_{\text{train}}, \sigma^2)$  (with the same  $\sigma^2$ ).

#### 5.2.4 GREEDY TOTAL UNCERTAINTY MINIMIZATION

While MAXDET satisfies (INF) and (DIV), it does not satisfy (REP) since it does not incorporate the pool set distribution. To fix this, Ash et al. (2021) propose BAIT. The regression version of BAIT tries to minimize the sum of the GP posterior variances on the training and pool set.<sup>3</sup> In other words, BAIT aims to minimize the acquisition function

$$a(\mathcal{X}_{\text{sel}}) := \sum_{\tilde{\mathbf{x}} \in \mathcal{X}_{\text{train}} \cup \mathcal{X}_{\text{pool}}} k_{\rightarrow\text{post}}(\mathcal{X}_{\text{sel}}, \sigma^2)(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}). \quad (23)$$

3. There is an “unregularized” version of BAIT that extends to classification, but we use the “regularized” version with  $\sigma^2 > 0$  here since it is more natural for GPs and avoids numerical issues.

In Section D.5, we show that Eq. (23) is equivalent to the original BAIT formulation. This is also known as (Bayesian) V-optimal design (Montgomery, 2017) and a similar method has been studied for NNs by Cohn (1996). Ash et al. (2021) propose two alternative methods for efficient approximate optimization of this acquisition function. The first one, which we call BAIT-F, is to simply use greedy selection as for the other methods in this section. The second alternative, which we call BAIT-FB, greedily selects  $2N_{\text{batch}}$  points (forward step) and then greedily removes  $N_{\text{batch}}$  points (backward step). In our framework, we can define BAIT-F by

$$\text{NEXTSAMPLE}(k, \mathcal{X}_{\text{sel}}, \mathcal{X}_{\text{rem}}) := \operatorname{argmin}_{\mathbf{x} \in \mathcal{X}_{\text{rem}}} \sum_{\tilde{\mathbf{x}} \in \mathcal{X}_{\text{train}} \cup \mathcal{X}_{\text{pool}}} k_{\rightarrow \text{post}}(\mathcal{X}_{\text{sel}} \cup \{\mathbf{x}\}, \sigma^2)(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}) .$$

Details on BAIT-F and BAIT-FB are given in Section D.5. Like for MAXDET, applying BAIT-F-TP or BAIT-FB-TP to a kernel  $k$  is equivalent to applying BAIT-F-P or BAIT-FB-P to  $k_{\rightarrow \text{post}}(\mathcal{X}_{\text{train}}, \sigma^2)$  (with the same  $\sigma^2$ ).

### 5.2.5 FRANK-WOLFE OPTIMIZATION

In order to make  $\mathcal{X}_{\text{batch}}$  representative of the pool set, Pinsler et al. (2019) suggest to choose  $\mathcal{X}_{\text{batch}}$  such that  $\sum_{\mathbf{x} \in \mathcal{X}_{\text{pool}}} \phi(\mathbf{x})$  is well-approximated by  $\sum_{\mathbf{x} \in \mathcal{X}_{\text{batch}}} w_{\mathbf{x}} \phi(\mathbf{x})$ , where  $w_{\mathbf{x}}$  are non-negative weights. Specifically, they propose to apply the Frank-Wolfe optimization algorithm to a corresponding optimization problem, which automatically selects elements of  $\mathcal{X}_{\text{batch}}$  iteratively. This can be seen as an attempt to represent the distribution of  $\mathcal{X}_{\text{pool}}$  with  $\mathcal{X}_{\text{batch}}$  by approximating the empirical kernel mean embedding  $N_{\text{pool}}^{-1} \sum_{\mathbf{x} \in \mathcal{X}_{\text{pool}}} k(\mathbf{x}, \cdot)$  using  $\mathcal{X}_{\text{batch}}$ . The corresponding selection method can be implemented in kernel space or feature space. Since the kernel space version scales quadratically with  $N_{\text{pool}}$ , Pinsler et al. (2019) use the feature space version for large pool sets. We also use the feature space version for our experiments and show the pseudocode in Section D.6. While the version by Pinsler et al. (2019) allows to select the same  $\mathbf{x} \in \mathcal{X}_{\text{pool}}$  multiple times, we prohibit this as it would allow to select smaller batches and thus prevent a fair comparison to other methods.

### 5.2.6 GREEDY DISTANCE MAXIMIZATION

A simple strategy to enforce the diversity of a set of points is to greedily select points with maximum distance to all previously selected points. The resulting algorithm has been frequently proposed in the literature under different names, see Section D.7. In our case, the kernel  $k$  gives rise to a distance measure  $d_k(\mathbf{x}, \tilde{\mathbf{x}})$  as in Eq. (13). With this distance measure, the MAXDIST selection method is specified by

$$\text{NEXTSAMPLE}(k, \mathcal{X}_{\text{sel}}, \mathcal{X}_{\text{rem}}) = \operatorname{argmax}_{\mathbf{x} \in \mathcal{X}_{\text{rem}}} \min_{\mathbf{x}' \in \mathcal{X}_{\text{sel}}} d_k(\mathbf{x}, \mathbf{x}') .$$

If the argmax is not unique, an arbitrary maximizer is chosen. If  $\mathcal{X}_{\text{sel}}$  is empty, we choose  $\operatorname{argmax}_{\mathbf{x} \in \mathcal{X}_{\text{rem}}} k(\mathbf{x}, \mathbf{x})$ .

The use of MAXDIST-TP with  $k_{\text{lin}}$  for BMAL has been suggested by Yu and Kim (2010), and with a kernel similar to  $k_{\text{ll}}$  for BMDAL by Sener and Savarese (2018) and Geifman and El-Yaniv (2017). Sener and Savarese (2018) also alternatively propose a more involved discrete optimization algorithm. In their experiments, MAXDIST yielded only slightly worse

results than the more involved optimization algorithm while being significantly faster and easier to implement. They note that the batch selected by MAXDIST is suboptimal with respect to a covering objective by at most a factor of two. In Section D.7, we show that a similar guarantee can be given when applying MAXDIST to a sketched approximation of the desired kernel. The use of dimensionality reduction for MAXDIST has also been analyzed by Eppstein et al. (2020). Inspired by the reasoning of Sener and Savarese (2018), we interpret the distances as uncertainty estimates: If both the optimal regression function  $f_*$  and the learned regression function  $f_{\theta_T}$  are  $L$ -Lipschitz with respect to  $d_k$ , and we have  $y_i = f_*(\mathbf{x}_i) = f_{\theta_T}(\mathbf{x}_i)$  on the training set, then we have the worst-case bound

$$|f_{\theta_T}(\mathbf{x}) - f_*(\mathbf{x})| \leq 2L \min_{\tilde{\mathbf{x}} \in \mathcal{X}_{\text{train}}} d_k(\mathbf{x}, \tilde{\mathbf{x}}). \quad (24)$$

Of course, the Lipschitz constant  $L$  might itself depend on  $\mathcal{X}_{\text{train}}$ , so this should only be interpreted as a crude heuristic. Wenzel et al. (2021) show that for kernel interpolation with Sobolev kernels, MAXDIST and MAXDET with  $\sigma^2 = 0$  yield asymptotically equivalent convergence rates.

### 5.2.7 $k$ -MEANS++ SEEDING

Similar to MAXDET, MAXDIST enforces (INF) and (DIV) but not (REP). To incorporate (REP), i.e., sample more points from regions with higher pool set density, we can view batch selection as a clustering problem: For example, if the distance-based uncertainty estimate in Eq. (24) holds, we could try to minimize the corresponding upper bound on the pool set MSE  $\frac{1}{N_{\text{pool}}} \sum_{\mathbf{x} \in \mathcal{X}_{\text{pool}}} |f_{\theta_T}(\mathbf{x}) - f_*(\mathbf{x})|^2$  after adding  $\mathcal{X}_{\text{batch}}$ :

$$\mathcal{X}_{\text{batch}} = \underset{\mathcal{X}_{\text{batch}} \subseteq \mathcal{X}_{\text{pool}}, |\mathcal{X}_{\text{batch}}| = N_{\text{batch}}}{\text{argmin}} \frac{1}{N_{\text{pool}}} \sum_{\mathbf{x} \in \mathcal{X}_{\text{pool}}} \min_{\tilde{\mathbf{x}} \in \mathcal{X}_{\text{mode}} \cup \mathcal{X}_{\text{batch}}} d_k(\mathbf{x}, \tilde{\mathbf{x}})^2. \quad (25)$$

This optimization problem is essentially the k-medoids problem (Kaufman and Rousseeuw, 1990), which combines the objective for the k-means clustering algorithm (Lloyd, 1982) with the constraint that the cluster centers must be chosen from within the data to be clustered. For large pool sets, common k-medoids algorithms can be computationally infeasible. An efficient approximate k-medoids solution can be computed using the seeding method of the k-means++ algorithm (Arthur and Vassilvitskii, 2007; Ostrovsky et al., 2006), which simply chooses the next batch element randomly via the distribution

$$\forall \mathbf{x} \in \mathcal{X}_{\text{rem}} : P(\text{NEXTSAMPLE}(k, \mathcal{X}_{\text{sel}}, \mathcal{X}_{\text{rem}}) = \mathbf{x}) = \frac{\min_{\tilde{\mathbf{x}} \in \mathcal{X}_{\text{sel}}} d_k(\mathbf{x}, \tilde{\mathbf{x}})^2}{\sum_{\mathbf{x}' \in \mathcal{X}_{\text{rem}}} \min_{\tilde{\mathbf{x}} \in \mathcal{X}_{\text{sel}}} d_k(\mathbf{x}', \tilde{\mathbf{x}})^2},$$

and if  $\mathcal{X}_{\text{sel}}$  is empty, it selects  $\text{NEXTSAMPLE}(k, \mathcal{X}_{\text{sel}}, \mathcal{X}_{\text{rem}})$  uniformly at random from  $\mathcal{X}_{\text{rem}}$ . We refer to the corresponding selection method as KMEANSPP. For the case  $\mathcal{X}_{\text{mode}} = \emptyset$ , Arthur and Vassilvitskii (2007) showed that with respect to the objective in Eq. (25), the batch selected by KMEANSPP is suboptimal by a factor of at most  $16 + 8 \log(N_{\text{batch}})$  in expectation. The use of KMEANSPP-P for BMDAL has been proposed in the so-called BADGE method by Ash et al. (2019). In contrast to our setting, BADGE is designed for classification and introduces an uncertainty estimate into  $k_{\parallel}$  not through a posterior transformation but through the influence of the softmax output layer on the magnitude of the gradients.

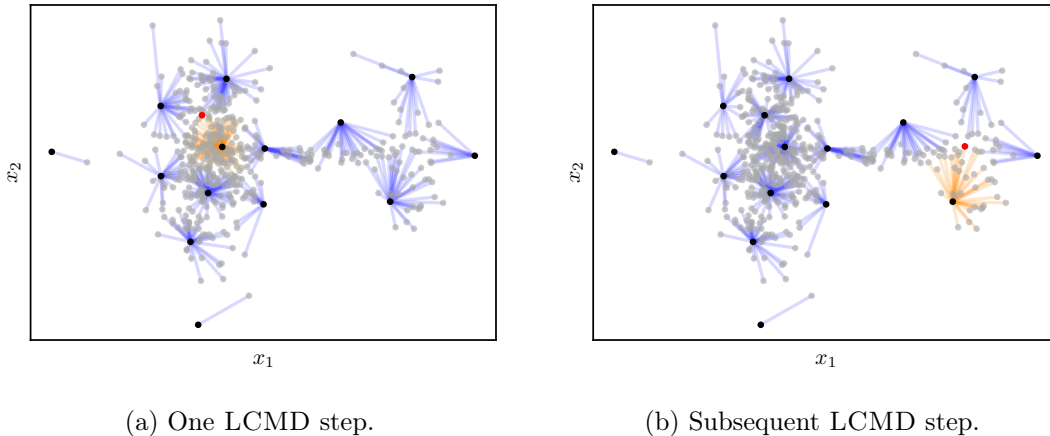


Figure 2: Two steps of LCMD selection for points  $\mathbf{x} \in \mathbb{R}^2$  with linear feature map  $\phi(\mathbf{x}) = \mathbf{x}$ . The black points represent the already selected points  $\mathcal{X}_{\text{sel}}$  and the gray points represent the remaining points  $\mathcal{X}_{\text{rem}}$ . The lines associate each remaining point to the closest selected point, forming clusters. The orange lines represent the cluster with the largest sum of squared distances. The red point, which is the remaining point with the largest distance to the cluster center within the orange cluster, is selected next. In the left plot, it can be seen that the smaller-radius cluster on the left is preferred over the larger-radius clusters on the right due to the higher point density on the left. After adding a point from the cluster, its size shrinks and a cluster on the right becomes dominant, which is shown in the right plot.

### 5.2.8 LARGEST CLUSTER MAXIMUM DISTANCE

As an alternative to the randomized KMEANSPP method, we propose a novel deterministic method that is inspired by the same objective (Eq. (25)). This method, which we call LCMD (largest cluster maximum distance), is visualized in Figure 2. Intuitively, we enforce (REP) by limiting the selection to the largest cluster, while we also enforce (DIV) by picking the maximum distance point within this cluster. LCMD can be formally defined as follows: We interpret points  $\tilde{\mathbf{x}} \in \mathcal{X}_{\text{sel}}$  as cluster centers. For each point  $\mathbf{x} \in \mathcal{X}_{\text{rem}}$ , we define its associated center  $c(\mathbf{x}) \in \mathcal{X}_{\text{sel}}$  as

$$c(\mathbf{x}) := \operatorname{argmin}_{\tilde{\mathbf{x}} \in \mathcal{X}_{\text{sel}}} d_k(\mathbf{x}, \tilde{\mathbf{x}}) .$$

If there are multiple minimizers, we pick an arbitrary one of them. Then, for each center  $\tilde{\mathbf{x}} \in \mathcal{X}_{\text{sel}}$ , we define the cluster size

$$s(\tilde{\mathbf{x}}) := \sum_{\mathbf{x} \in \mathcal{X}_{\text{rem}}: c(\mathbf{x}) = \tilde{\mathbf{x}}} d_k(\mathbf{x}, \tilde{\mathbf{x}})^2 .$$

We then pick the maximum-distance point from the cluster with maximum size:

$$\text{NEXTSAMPLE}(k, \mathcal{X}_{\text{sel}}, \mathcal{X}_{\text{rem}}) = \operatorname{argmax}_{\mathbf{x} \in \mathcal{X}_{\text{rem}}: s(c(\mathbf{x})) = \max_{\tilde{\mathbf{x}} \in \mathcal{X}_{\text{sel}}} s(\tilde{\mathbf{x}})} d_k(\mathbf{x}, c(\mathbf{x})) .$$



As for MAXDIST, if  $\mathcal{X}_{\text{sel}}$  is empty, we choose  $\operatorname{argmax}_{\mathbf{x} \in \mathcal{X}_{\text{rem}}} k(\mathbf{x}, \mathbf{x})$  instead. If the selection of pool points should be adapted to the distribution of another set  $\mathcal{X}$  instead of  $\mathcal{X}_{\text{pool}}$ , one may simply compute the cluster sizes based on  $\mathcal{X}$  instead. Importantly, like KMEANSPP but unlike some other k-medoids methods, LCMD can be implemented with a runtime complexity that is linear in  $N_{\text{pool}}$ , as discussed in Section D.9.

### 5.2.9 OTHER OPTIONS

As mentioned above, MAXDIST can be interpreted as a greedy optimization algorithm for a covering objective that is NP-hard to (approximately) optimize (Gonzalez, 1985; Feder and Greene, 1988), but for which other approximate optimization algorithms have been proposed (Sener and Savarese, 2018). Similarly, MAXDET-P attempts to greedily maximize  $\det(k(\mathcal{X}_{\text{batch}}, \mathcal{X}_{\text{batch}}) + \sigma^2 \mathbf{I})$ , which is NP-hard to (approximately) optimize (Civril and Magdon-Ismail, 2013), but for which other approximate optimization algorithms have been proposed (Biyik et al., 2019). We do not investigate these advanced optimization algorithms here as they come with greatly increased runtime cost, and MAXDET and MAXDIST already enjoy some approximation guarantees, as discussed in Section D.

Yu and Kim (2010), Wu (2018), and Zhdanov (2019) suggest less scalable clustering-based approaches for BM(D)AL. Alternatively, it might be interesting to try the greedy  $k$ -means++ algorithm (Celebi et al., 2013), which provides a slightly less efficient alternative to the  $k$ -means++ algorithm.

Caselton and Zidek (1984) propose to optimize mutual information between the batch samples and the remaining pool samples, which is analyzed for GPs by Krause et al. (2008), but does not scale well with the pool set size for GPs, at least in a general kernel-space formulation. Another option is to remove the non-negativity constraint on the weights  $w_{\mathbf{x}}$  used in FRANKWOLFE. This setting is also treated in a generalized fashion in Santin et al. (2021). An investigation of this method is left to future work.

## 5.3 Discussion

When considering the design criteria from Section 2.2, we can say that MAXDIAG only satisfies (INF), while MAXDET and MAXDIST also satisfy (DIV). Arguably, BAIT, KMEANSPP, and LCMD satisfy all three properties (INF), (DIV) and (REP). The FRANKWOLFE method is only designed to satisfy (REP), based on which one could argue that (INF) and (DIV) are also satisfied to some extent.

In terms of runtime complexity, as can be seen in Table 3, all considered selection methods are well-behaved for moderate feature space dimension  $d_{\text{feat}}$ . When considering kernels such as  $k_{\text{grad}}$ , whose evaluation is tractable despite having very high feature space dimension, distance-based selection methods are still efficient while BAIT and FRANKWOLFE can become intractable for large pool set sizes, and MAXDET exhibits worse scaling with respect to  $N_{\text{batch}}$ . Moreover, if  $\phi(\mathcal{X}_{\text{train}})$  has full rank and  $d_{\text{feat}} \leq N_{\text{train}}$ , it follows from Eq. (10) that in the limit  $\sigma^2 \rightarrow 0$ , the GP posterior uncertainty becomes zero everywhere. Hence, Bayesian posterior-based methods like MAXDET, BAIT and FRANKWOLFE might require  $d_{\text{feat}} \gtrsim N_{\text{train}}$  for good performance, which in turn deteriorates their runtime.

In the non-batch active learning setting, that is, for  $N_{\text{batch}} = 1$ , some selection methods become equivalent: LCMD-P, MAXDIST-P, and MAXDET-P become equivalent to MAX-

DIAG-P; moreover, KMEANSPP-P becomes equivalent to RANDOM. This suggests that for  $N_{\text{batch}} = 1$ , TP-mode is necessary for KMEANSPP and LCMD to ensure (REP).

## 6. Experiments

To evaluate a variety of combinations of kernels, kernel transformations, and selection methods, we introduce a new open-source benchmark for BMDAL for regression. Our benchmark uses 15 large tabular regression data sets, with input dimensions ranging between two and 379, that are selected mostly from the UCI and OpenML repositories, cf. the detailed description in Section E.1. The initial pool set size  $N_{\text{pool}}$  for these data sets lies between 31335 and 198720. As an NN model, we use a three-layer fully connected NN with 512 neurons in both hidden layers, parameterized as in Section 2.1. We train the NN using Adam (Kingma and Ba, 2015) for 256 epochs with batch size 256, using early stopping based on a validation set. Results shown here are for the ReLU activation function, but we also re-ran most of our experiments for the SiLU (a.k.a. swish) activation (Elfwing et al., 2018), and unless indicated otherwise, our insights discussed below apply to results for both activation functions. We manually optimized the parameters  $\sigma_w, \sigma_b$  and the learning rate separately for both activation functions to optimize the average logarithmic RMSE of RANDOM selection. Details on the NN architecture and training are described in Section E.3. For the hyperparameter  $\sigma^2$ , which occurs in MAXDET and various posterior-based transformations, we found that smaller values typically yield better average results but may cause numerical instabilities. As a compromise, we chose  $\sigma^2 = 10^{-6}$  in our experiments and use 64-bit floats for computations involving  $\sigma^2$ .

In our evaluation, we start with  $N_{\text{train}} = 256$  and then acquire 16 batches with  $N_{\text{batch}} = 256$  samples each using the respective BMAL method. We repeat this 20 times with different seeds for NN initialization and different splits of the data into training, validation, pool, and test sets. We measure the mean absolute error (MAE), root mean squared error (RMSE), 95% and 99% quantiles, and the maximum error (MAXE) on the test set after each BMAL step. For each of those five error metrics, we average the logarithms of the metric over the 20 repetitions, and, depending on the experiment, over the 16 steps and/or the 15 data sets. Note that a difference of  $\delta$  between two logarithmic values corresponds to a ratio  $e^\delta \approx 1 + \delta$  between the values; for example, a reduction by  $\delta = 0.1$  corresponds to a reduction of the geometric mean error by about 10%. Our most important metric is the RMSE, but we will also put some focus on MAXE since it can be interpreted as a measure of robustness to distribution shifts. Generally, RMSE is more affected by rare but large errors than MAE, while the quantiles and MAXE exclusively focus on rare but large errors.

In the following, we will discuss some of the benchmark results. More detailed results can be found in Section E.4.

### 6.1 Comparison to Existing Methods

Based on our detailed evaluation in Table E.5 and Table E.6, we propose a new BMDAL algorithm as the combination of the LCMD-TP selection method and the kernel  $k_{\text{grad} \rightarrow \text{sketch}(512)}$ . Figure 1 and Table 5 show that our proposed combination clearly outperforms other methods from the literature in terms of averaged logarithmic RMSE over our benchmark data

sets and random splits. We incorporate the methods from the literature into our framework as shown in Table 5, which involves the following modifications:

- The BALD (Houlsby et al., 2011) and BatchBALD (Kirsch et al., 2019) acquisition functions are applied to a last-layer Gaussian Process model.
- For BAIT, we rescale  $k_{\parallel}$  based on the training set before applying the posterior transformation, see Section 4.2.1, and we apply regularization by using a small  $\sigma^2 > 0$ .
- For ACS-FW (Pinsler et al., 2019), we use the FRANKWOLFE selection method with  $k_{\parallel \rightarrow \text{acs-rf-hyper}(512)}$ . Compared to the experiments by Pinsler et al. (2019), there are several differences: First, we do not permit FRANKWOLFE to select smaller batches by selecting the same point multiple times. Second, we use 512 random features instead of 10. Third, our acs-rf-hyper transformation first rescales  $k_{\parallel}$  based on the training set, which we found to improve performance. Fourth, our  $k_{\parallel}$  kernel incorporates the last-layer bias and not only the weights.
- By Core-Set, we refer to the k-center-greedy method of Sener and Savarese (2018) applied to  $k_{\parallel}$ , which is also equivalent to FF-Active (Geifman and El-Yaniv, 2017).
- For BADGE (Ash et al., 2019), which originally incorporates uncertainties into  $\phi_{\parallel}$  through softmax gradients, we use  $\phi_{\parallel \rightarrow \mathcal{X}_{\text{train}}}$  instead of  $\phi_{\parallel}$ .

As argued in Section 2.2, we do not compare to methods that require training with ensembles (Krogh and Vedelsby, 1994), since ensembles are more expensive to train and these methods are typically designed for non-batch active learning. Moreover, we do not compare to methods that require training with Dropout (Tsybalov et al., 2018) or custom loss functions (Ranganathan et al., 2020), since these methods can change the error for the underlying NN, which makes them difficult to compare fairly and more inconvenient to use.

## 6.2 Evaluated Combinations

Our framework allows us to obtain a vast number of BMDAL algorithms via combinations of base kernels, kernel transformations, and the P and TP modes of different selection methods. Table E.5 and Table E.6 in Section E.4 show a large number of such combinations for ReLU and SiLU activations, respectively. These combinations have been selected according to the following principles:

- Kernels for P-mode selection use posterior-based transformations, while kernels for TP-mode selection do not (see Section 5.1).
- Sketching and random features always use 512 target features. Similar to the hidden layer size of 512, this number has been selected to be a bit larger than the usually employed  $N_{\text{batch}} = 256$ . Note that due to the bias in the last layer,  $k_{\parallel}$  has a 513-dimensional feature space.
- FRANKWOLFE is only run in P mode (as proposed in Pinsler et al. (2019)), since in TP mode, kernel mean embeddings in 512-dimensional feature space would be approximated using more than 512 samples. Note that Pinsler et al. (2019) only use 10 instead of 512 random features in their experiments, leading to a worse approximation.
- Due to the equivalence between P mode with posterior transformation and TP mode without posterior transformation mentioned in Section 5.2.3, BAIT is always run in P mode, and MAXDET is mostly run in P mode except for  $k_{\text{grad}}$  and  $k_{\text{nngp}}$  due to their high-dimensional feature space.

BMDAL method	Selection method	Kernel	mean log RMSE ( $\downarrow$ )	
			ReLU	SiLU
Supervised learning	RANDOM	—	-1.401	-1.406
BALD (Houlsby et al., 2011) with last-layer GP	MAXDIAG	$k_{\parallel \rightarrow \mathcal{X}_{\text{train}}}$	-1.285	-1.300
BatchBALD (Kirsch et al., 2019) with last-layer GP	MAXDET-P	$k_{\parallel \rightarrow \mathcal{X}_{\text{train}}}$	-1.463	-1.467
BAIT (Ash et al., 2021)	BAIT-FB-P	$k_{\parallel \rightarrow \mathcal{X}_{\text{train}}}$	-1.541	-1.522
ACS-FW (Pinsler et al., 2019)	FRANKWOLFE-P	$k_{\parallel \rightarrow \text{acs-rf-hyper}(512)}$	-1.439	-1.437
Core-Set* (Sener and Savarese, 2018), FF-Active (Geifman and El-Yaniv, 2017)	MAXDIST-TP	$k_{\parallel}$	-1.491	-1.515
BADGE (Ash et al., 2019) with last-layer GP uncertainty	KMEANSPP-P	$k_{\parallel \rightarrow \mathcal{X}_{\text{train}}}$	-1.530	-1.484
Ours	LCMD-TP	$k_{\text{grad} \rightarrow \text{sketch}(512)}$	<b>-1.590</b>	<b>-1.597</b>

\* This refers to their simpler k-center-greedy selection method.

Table 5: Comparison of our BMDAL method against other methods from the literature (cf. Table 4). The mean log RMSE is averaged over all data sets, repetitions, and BMAL steps for the respective experiments with ReLU or SiLU activation function. We make small adjustments to the literature methods as described in Section 6.

In general, we observe the following trends in our results across selection methods:

- The network-dependent base kernels  $k_{\parallel}$  and  $k_{\text{grad}}$  clearly outperform the network-independent base kernels  $k_{\text{lin}}$  and  $k_{\text{nnngp}}$  across different selection methods, modes and kernel transformations.
- Out of the network-dependent base kernels,  $k_{\text{grad}}$  typically outperforms  $k_{\parallel}$ , at least for NN hyperparameters optimized for RANDOM (cf. Section E.3). It should be noted that in our ReLU experiments, these optimized hyperparameters typically lead to many dead neurons in the last hidden layer, which may affect  $k_{\parallel}$  by reducing the effective feature space dimension.<sup>4</sup> We define the effective feature space dimension of the pool set for a kernel  $k$  as

$$d_{\text{eff}} := \frac{\text{tr}(k(\mathcal{X}_{\text{pool}}, \mathcal{X}_{\text{pool}}))}{\|k(\mathcal{X}_{\text{pool}}, \mathcal{X}_{\text{pool}})\|_2} = \frac{\lambda_1 + \dots + \lambda_{d_{\text{feat}}}}{\lambda_1},$$

where  $\lambda_1 \geq \dots \geq \lambda_{d_{\text{feat}}}$  are the eigenvalues of the feature covariance matrix

$$\phi(\mathcal{X}_{\text{pool}})^\top \phi(\mathcal{X}_{\text{pool}}) \in \mathbb{R}^{d_{\text{feat}} \times d_{\text{feat}}}.$$

4. In the extreme case where all neurons in the last hidden layer are dead, the network-dependent base kernels become degenerate, which can cause numerical problems in selection methods. Once a selection method suggests an invalid (e.g. already selected) sample for the batch, we fill up the rest of the batch with random samples. Out of the 597900 BMDAL steps in our ReLU experiments, such invalid samples were suggested in just 4 steps in total.

With this definition,  $d_{\text{eff}}$  is indeed typically much larger for  $k_{\text{grad} \rightarrow \text{sketch}(512)}$  than for  $k_{\parallel}$  in our experiments.<sup>5</sup> Another difference is that, for the ReLU activation function,  $k_{\text{grad}}$  is discontinuous while  $k_{\parallel}$  is not.

- For  $k_{\text{grad}}$ , applying sketching does not strongly affect the resulting accuracy while leading to considerably faster runtimes.
- When evaluating the use of ensembled NN kernels, we want to differentiate between the effect of ensembling on the accuracy of supervised learning and the effect of ensembling on the quality of the selected batches  $\mathcal{X}_{\text{batch}}$ . To eliminate the former effect, we only consider the averaged accuracies of the individual ensemble members and not the accuracy of their averaged predictions. With this method of evaluation, we find that ensembling of network-dependent kernels only leads to small improvements in accuracy, at least for the ensembling configurations we tested. This is in contrast to other papers where the uncertainty of the ensemble predictions turned out to be more beneficial (Beluch et al., 2018; Pop and Fulop, 2018). Perhaps ensembling is less useful in our case because our non-ensembled kernels already provide good uncertainty measures.
- Out of the acs-grad, acs-rf and acs-rf-hyper transformations, acs-rf often performs best, except for FRANKWOLFE-P with base kernel  $k_{\text{grad}}$ , where acs-rf-hyper performs best.
- In contrast to Ash et al. (2021), we find that BAIT-FB does not perform better than BAIT-F.
- The relative gains for BMDAL methods compared to RANDOM selection are typically largest on metrics such as MAXE or 99% quantile, and worst on MAE.
- All investigated BMDAL methods only take a few seconds to select a batch in our experiments on our NVIDIA RTX 3090 GPUs (cf. Section E), which is typically faster than the time for training the corresponding NN. Hence, we expect all investigated BMDAL methods to be much faster than the time for labeling in most scenarios where BMDAL is desirable. Note that the runtime of P-mode selection methods is comparable to those of TP-mode selection methods only because we typically run P-mode selection with 64-bit floats to avoid numerical issues for posteriors. TP-mode selection methods need to consider  $N_{\text{train}} + N_{\text{batch}}$  instead of  $N_{\text{batch}}$  selected points, which can be significantly slower than P-mode if  $N_{\text{train}} \gg N_{\text{batch}}$ . Especially for TP-mode selection methods, it may therefore be desirable to let  $N_{\text{batch}}$  grow proportionally to  $N_{\text{train}}$ .

### 6.3 Best Kernels and Modes for each Selection Method

Our base kernels, kernel transformations, and selection modes yield numerous ways to apply each selection method. To compare selection methods, we choose for each selection method the best-performing combination according to the averaged logarithmic RMSE, excluding kernels with ensembling and  $k_{\text{grad}}$  without sketching since they considerably increase computational cost while providing comparable accuracy to their more efficient counterparts. The selected combinations are shown in Table 6. Compared to the literature methods in

---

5. Specifically, averaged over all corresponding ReLU experiments with  $N_{\text{batch}} = 256$  and over all BMAL steps,  $k_{\text{grad} \rightarrow \text{sketch}(512)}$  leads to an average  $d_{\text{eff}}$  of about 5.5, while  $k_{\parallel}$  leads to an average  $d_{\text{eff}}$  of about 1.7. On corresponding BMAL steps, the effective dimension is larger for  $k_{\text{grad} \rightarrow \text{sketch}(512)}$  about 95% of the time. For SiLU, the results are slightly less extreme, with effective dimensions of 4 and 2.3, and the effective dimension of  $k_{\text{grad} \rightarrow \text{sketch}(512)}$  being larger about 90% of the time.

Selection method	Sel. mode	Selected kernel	mean log RMSE	Avg. time [s]
RANDOM	—	—	-1.401	0.001
MAXDIAG	—	$k_{\text{grad} \rightarrow \text{sketch}(512) \rightarrow \text{acs-rf}(512)}$	-1.370	0.650
MAXDET	P	$k_{\text{grad} \rightarrow \text{sketch}(512) \rightarrow \mathcal{X}_{\text{train}}}$	-1.512	0.770
BAIT	F-P	$k_{\text{grad} \rightarrow \text{sketch}(512) \rightarrow \mathcal{X}_{\text{train}}}$	-1.585	1.508
FRANKWOLFE	P	$k_{\text{grad} \rightarrow \text{sketch}(512) \rightarrow \text{acs-rf-hyper}(512)}$	-1.542	0.823
MAXDIST	P	$k_{\text{grad} \rightarrow \text{sketch}(512) \rightarrow \mathcal{X}_{\text{train}}}$	-1.514	0.713
KMEANSPP	P	$k_{\text{grad} \rightarrow \text{sketch}(512) \rightarrow \text{acs-rf}(512)}$	-1.569	0.836
LCMD	TP	$k_{\text{grad} \rightarrow \text{sketch}(512)}$	<b>-1.590</b>	0.981

Table 6: Selected kernels and modes per selection method that are shown in our plots. The mean log RMSE is averaged over all data sets, repetitions, and BMAL steps. The average time for the batch selection is averaged over all data sets and BMAL steps, measured at one repetition with only one process running on each NVIDIA RTX 3090 GPU. An overview of all results can be found in Table E.5.

Table 5, we see that optimizing the kernel and mode can yield a considerable difference in performance. Note that the relative performance between the configurations for SiLU is slightly different, as can be seen in Table E.6. If we selected combinations according to the SiLU results, the combinations for MAXDIST and KMEANSPP in Table 6 would use TP-mode and  $k_{\text{grad} \rightarrow \text{sketch}(512)}$  instead. When considering only kernels based on  $k_{\parallel}$ , the results from Table E.5 and Table E.6 show that a comparison of selection methods would look qualitatively similar.

#### 6.4 Comparison of Selection Methods

We compare the selected configurations from Table 6 in several aspects. Figure 3 shows the evolution of the mean logarithmic MAE, RMSE, MAXE, 95% quantile, and 99% quantile over the BMAL steps, for a batch size of  $N_{\text{batch}} = 256$ . This demonstrates that the best considered BMDAL methods can match the average performance of RANDOM selection with about half of the samples for RMSE, and even fewer samples for MAXE. On individual data sets, this may differ, as shown in Figure 4. From this figure, it is apparent that when considering RMSE, LCMD-TP outperforms other methods not only in terms of average performance but across the majority of the data sets. Specifically, Table E.3 shows that LCMD-TP matches or exceeds the performance of the other selected BMDAL methods on 8 out of the 15 data sets in terms of RMSE. Figure 4 also shows that the selected MAXDET-P and MAXDIST-P configurations yield very similar performance on all data sets.

Figure 5 shows the influence of the chosen batch size  $N_{\text{batch}}$  on the final performance at  $N_{\text{train}} = 4352$  training samples. As expected, the naive active learning scheme MAXDIAG is particularly sensitive to the batch size. The other selection methods are less sensitive to changes in  $N_{\text{batch}}$  and exhibit almost no degradation in performance up to  $N_{\text{batch}} = 1024$ . Note that this “threshold” might depend on the initial and final training set sizes as well as the feature-space dimension  $d_{\text{feat}}$ .

Overall, the discussed figures and the detailed results in Table E.5 show that LCMD-TP performs best in terms of MAE, RMSE, and 95% quantile, followed by BAIT-F-P and KMEANSPP. For MAXE, MAXDIST, MAXDET and BAIT-F-P exhibit the best performances. Since MAXDET and MAXDIST are motivated by worst-case considerations, it is not

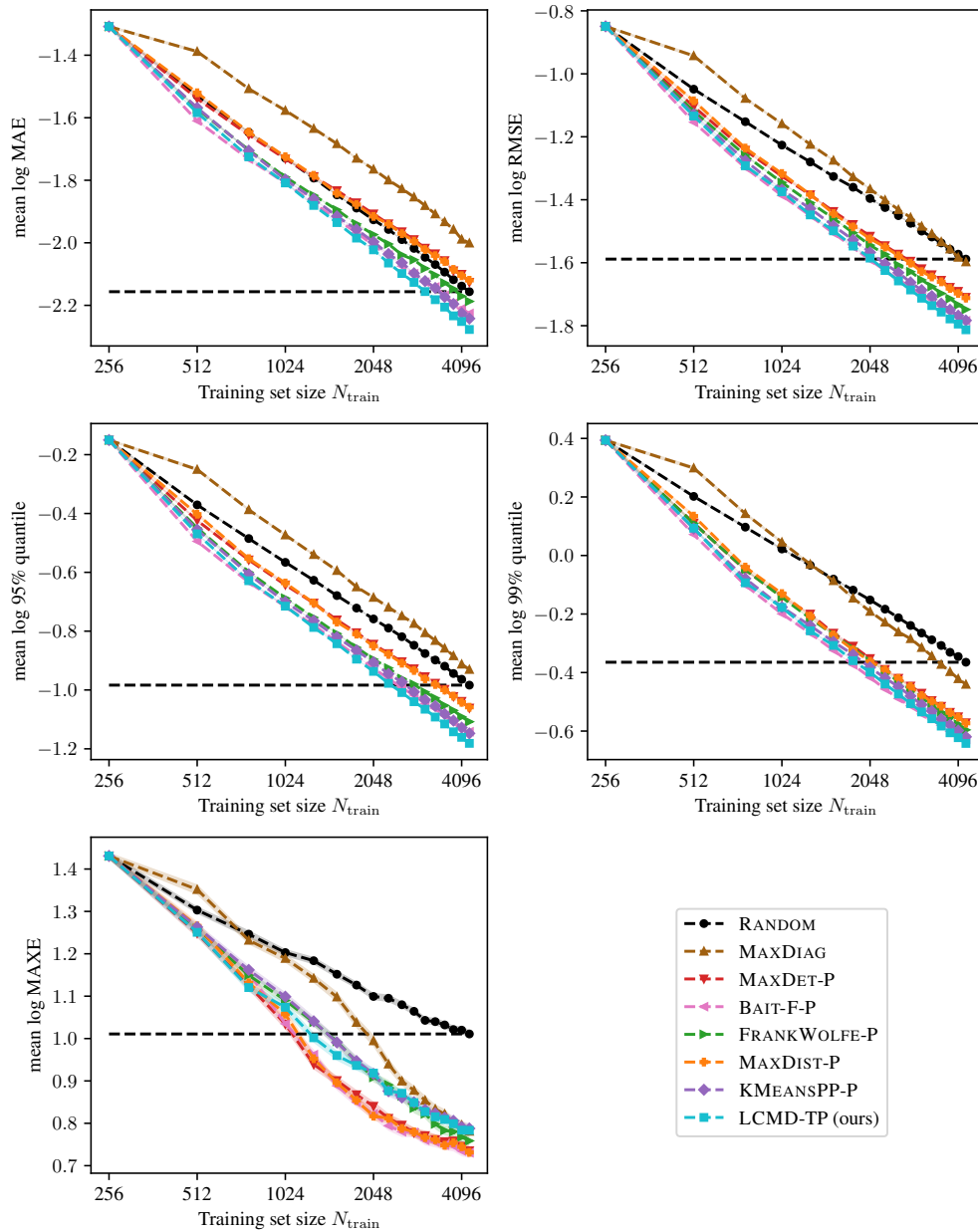


Figure 3: This figure shows how fast the errors decrease during BMAL for different selection methods and their corresponding kernels from Table 6. Specifically, for each of the five error metrics, the corresponding plot shows the logarithmic error metric between each BMAL step for  $N_{\text{batch}} = 256$ , averaged over all repetitions and data sets. The performance of RANDOM can be interpreted as the performance of supervised learning without active learning. The black horizontal dashed line corresponds to the final performance of RANDOM at  $N_{\text{train}} = 4352$ . The shaded area, which is nearly invisible for all metrics except MAXE, corresponds to one estimated standard deviation of the mean estimator, cf. Section E.4.

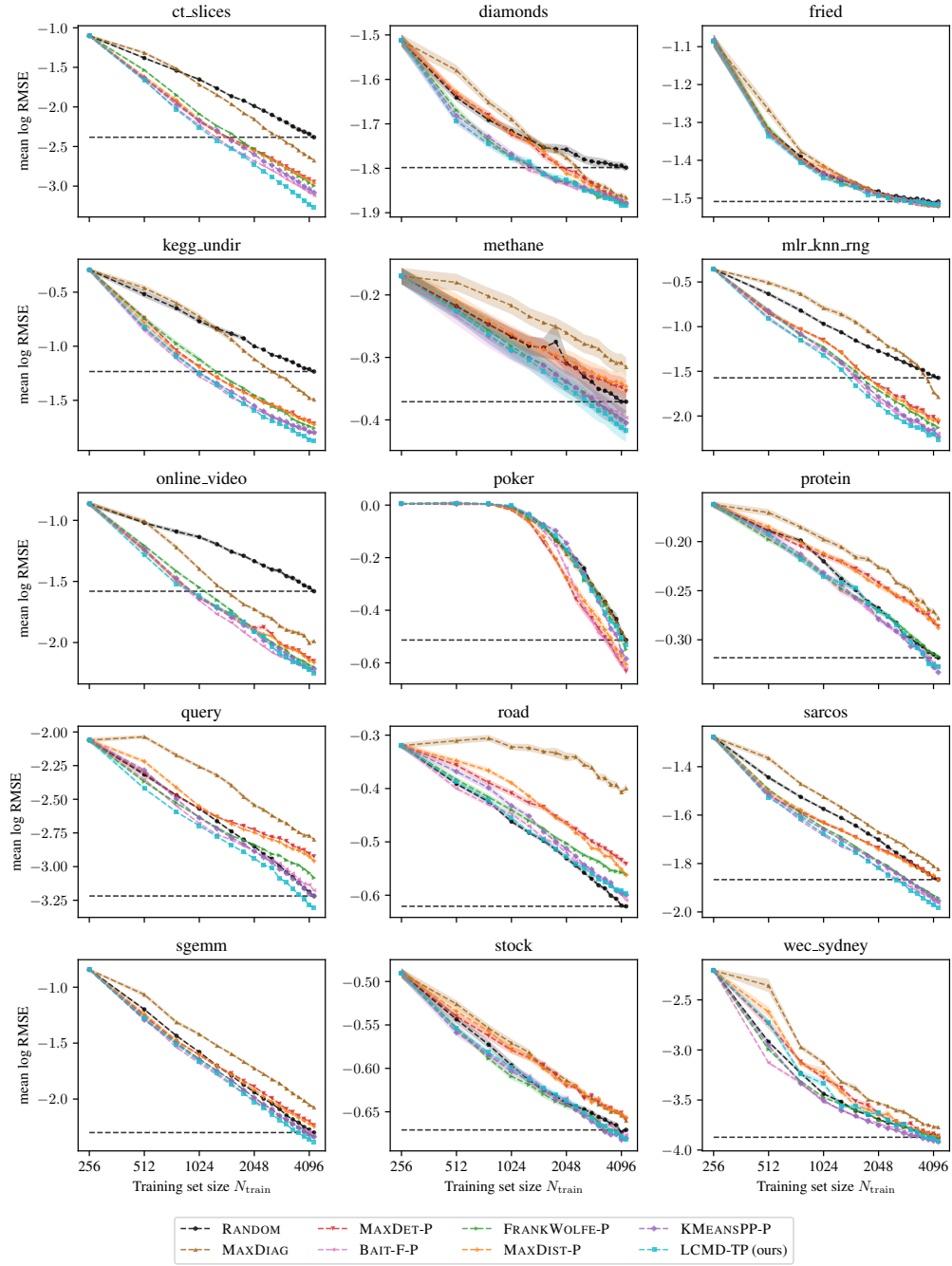


Figure 4: This figure shows how fast the RMSE decreases during BMAL on the individual benchmark data sets for different selection methods and their corresponding kernels from Table 6. Specifically, the plots above show the logarithmic RMSE between each BMAL step for  $N_{\text{batch}} = 256$ , averaged over all repetitions. The black horizontal dashed line corresponds to the final performance of RANDOM at  $N_{\text{train}} = 4352$ . The shaded area corresponds to one estimated standard deviation of the mean estimator, cf. Section E.4.



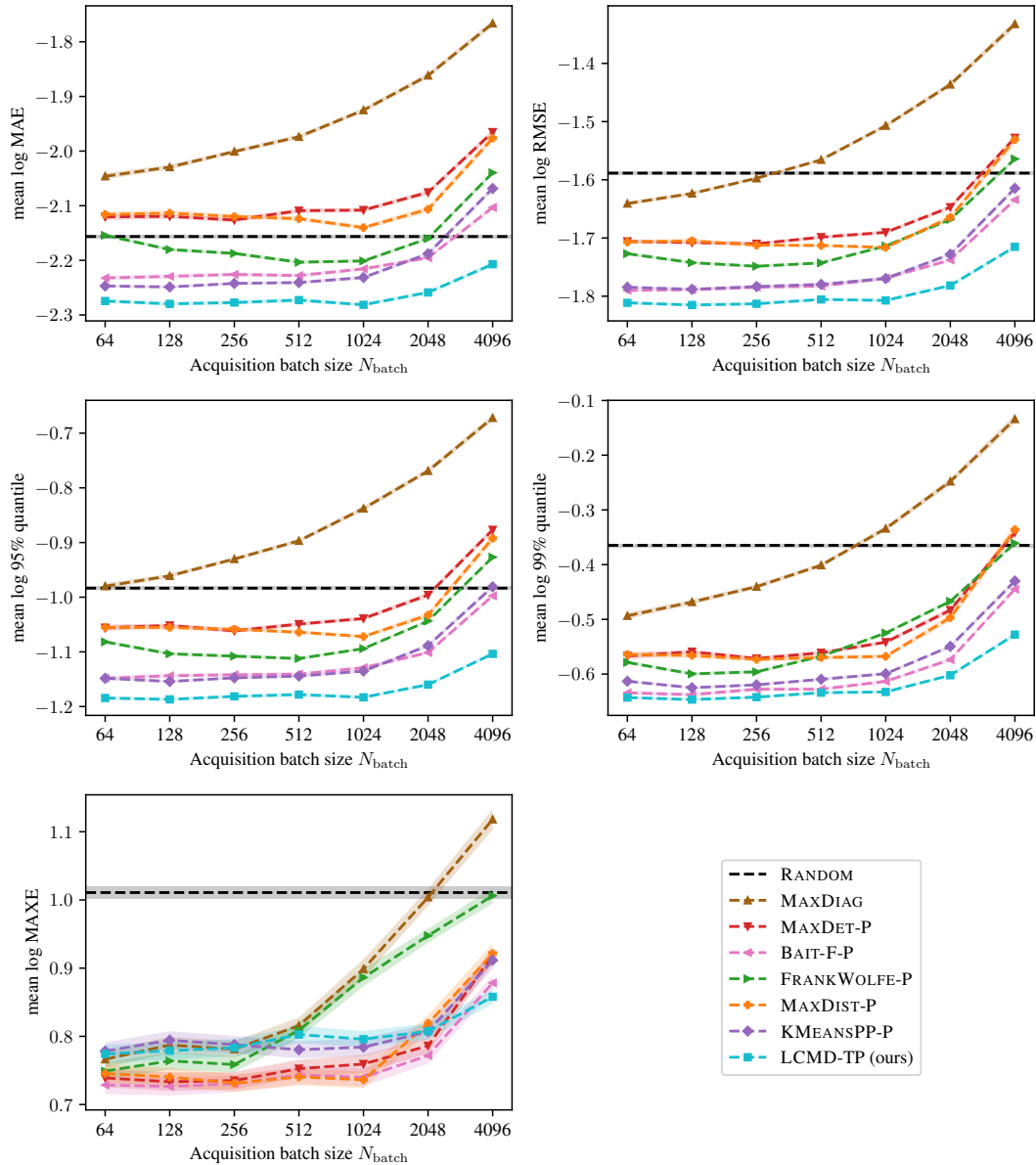


Figure 5: This figure shows how much the final accuracy of different BMDAL methods deteriorates when fewer BMAL steps with larger batch sizes are used. Specifically, we use different selection methods with the corresponding kernels from Table 6, starting with  $N_{\text{train}} = 256$  and then performing  $2^m$  BMAL steps with batch size  $N_{\text{batch}} = 2^{12-m}$  for  $m \in \{0, \dots, 6\}$ , such that the final training set size is 4352 in each case. For each of the five error metrics, the corresponding plot shows the final logarithmic error metric, averaged over all data sets and repetitions. Note that the performance of RANDOM selection does not depend on  $N_{\text{batch}}$  but only on the final training set size, hence it is shown as a constant line here. The shaded area, which is nearly invisible for all metrics except MAXE, corresponds to one estimated standard deviation of the mean estimator, cf. Section E.4.

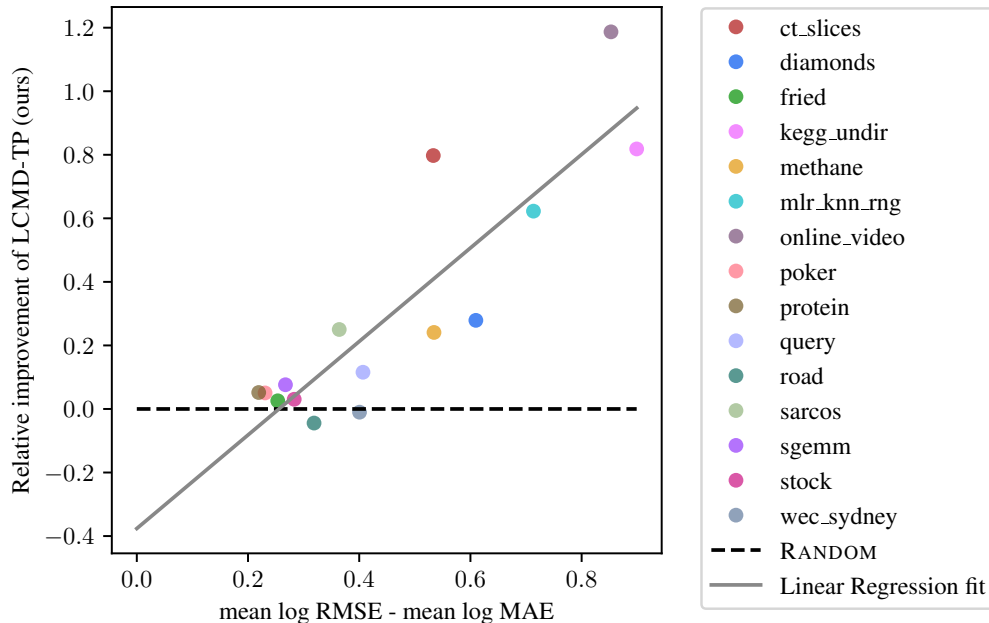


Figure 6: This figure shows, for each data set, the improvement in sample efficiency of LCMD-TP over RANDOM (on the  $y$ -axis) versus the variation of the error distribution (on the  $x$ -axis). The variation is measured as mean log RMSE – mean log MAE on the initial training set ( $N_{\text{train}} = 256$ ). The improvement in sample efficiency is measured by  $\frac{\text{mean log RMSE}(\text{LCMD-TP}) - \text{mean log RMSE}(\text{RANDOM})}{\text{mean log RMSE}(\text{RANDOM}) - (\text{mean log RMSE at } N_{\text{train}}=256)}$ . Here, the means are taken over all 20 repetitions and, unless indicated otherwise, over the trained networks after each of the 16 BMDAL steps. The Pearson correlation coefficient for the plotted data is  $R \approx 0.88$ .

surprising that they perform well on MAXE, while the strong performance of BAIT-F-P on MAXE is unexpected. Moreover, it is perhaps surprising that in Figure 3, the relative performances for the 99% quantile are arguably more similar to the RMSE performances than to the MAXE performances. For the 95% quantile, the relative performances for MAXDET, MAXDIST and MAXDIAG are even worse than for the RMSE. Thus, the use of MAXDIST instead of LCMD may only be advisable if one expects strong distribution shifts between pool and test sets.

### 6.5 When should BMDAL be Applied?

While LCMD-TP achieves excellent average performance, its benefits over RANDOM selection vary strongly between data sets, as is evident from Figure 4. Overall, LCMD-TP with  $k_{\text{grad} \rightarrow \text{sketch}(512)}$  outperforms random selection in terms of RMSE on 13 out of the 15 data sets and barely performs worse on the other two. Nonetheless, an *a priori* estimate of the benefits of LCMD-TP over RANDOM selection on individual data sets could be useful to inform practitioners on whether they should be interested in applying BMDAL. Figure 6 shows that on our 15 benchmark data sets, the variation of test errors after training on the

initial training set ( $N_{\text{train}} = 256$ ), measured by the quotient  $\frac{\text{RMSE}}{\text{MAE}}$ , is strongly correlated with the improvement in sample efficiency through LCMD-TP over RANDOM. In other words, the larger  $\frac{\text{RMSE}}{\text{MAE}}$  on the initial training set, the more benefit we can expect from LCMD-TP with  $k_{\text{grad} \rightarrow \text{sketch}(512)}$  over random selection.

## 7. Conclusion

In this paper, we introduced a framework to compose BMDAL algorithms out of base kernels, kernel transformations, and selection methods. We then evaluated different combinations of these components on a new benchmark consisting of 15 large tabular regression data sets. In our benchmark results, for all considered selection methods, replacing the wide-spread last-layer kernel  $k_{11}$  by a sketched finite-width neural tangent kernel  $k_{\text{grad} \rightarrow \text{sketch}(p)}$  leads to accuracy improvements at similar runtime and memory cost. Moreover, our novel LCMD selection method sets new state-of-the-art results in our benchmark in terms of RMSE and MAE.

### 7.1 Limitations

The BMDAL methods in our framework are very attractive for practitioners using NNs for regression since they are scalable to large data sets and can be applied to a wide variety of NN architectures and training methods without requiring modifications to the NN. However, while our benchmark contains many large data sets, it cannot cover all possible application scenarios that the considered BMDAL methods could be applied to. For example, it is unclear whether our insights can be transferred to applications like drug discovery (Mehrjou et al., 2021) or atomistic ML (Zaverkin and Kästner, 2021), where other types of data and other NNs are employed. Even in the tabular data setting, the relevance of our results for smaller data sets or recently proposed NN architectures (e.g. Gorishniy et al., 2021; Somepalli et al., 2021; Kadra et al., 2021) is unclear. Moreover, the current benchmark does not involve distribution shifts between pool and test data, which would be interesting for some practical applications.

### 7.2 Remaining Questions

Our results give rise to some interesting questions for future research, of which we list some in the following: Can  $k_{\text{grad}}$  be adapted to incorporate effects of optimizers such as Adam? How can it be efficiently evaluated and sketched for other types of layers? How can we decide which method to use on a data set, beyond just using the one with the best average performance across the benchmark? Are there some characteristics of data sets or training setups that can be used to predict which method will perform best? Are clustering-based methods like LCMD-TP also superior to other methods for non-batch AL? How much better performance could be attained if the methods had access to the pool labels, and what kinds of batches would such a method select? Would it have similar properties as Zhou et al. (2021) found for classification? How can our framework be generalized to classification or multi-output regression?

Our framework is formulated for the pool-based AL setting, where samples should be selected from a pool of unlabeled samples. By using a different kind of selection methods

with the same kernels and kernel transformations, our framework could be adapted to the streaming AL setting, where unlabeled samples arrive sequentially and one has to decide immediately whether to label them or not. For the membership-query AL setting, where unlabeled samples can be chosen arbitrarily, the situation is more difficult: Since the feature map  $\phi$  is typically not surjective, samples cannot be chosen in the feature space directly, and a direct choice in the input space might require differentiating the kernel for efficient optimization. Nonetheless, extending our methods to other AL settings could be an interesting avenue for further research.

## Acknowledgments

We want to thank Philipp Hennig, Tizian Wenzel, Daniel Winkle, Benjamin Unger, and Paul Bürkner for helpful comments. Funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany’s Excellence Strategy - EXC 2075 – 390740016. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting David Holzmüller. Viktor Zaverkin acknowledges the financial support received in the form of a Ph.D. scholarship from the Studienstiftung des Deutschen Volkes (German National Academic Foundation).

## Appendix A. Overview

The appendix structure mirrors the structure of the main paper, providing more details on the corresponding sections of the main paper: We discuss further details on base kernels in Section B, on kernel transformations in Section C, and on selection methods in Section D. The latter section includes efficiency-focused pseudocode as well as discussions on relations to the literature. Finally, we provide details on the setup and results of our experiments in Section E.

## Appendix B. Details on Base Kernels

In the following, we will provide details for the infinite-width NNGP kernel.

### B.1 NNGP Kernel

For the fully-connected NN model considered in Section 2.1 with a ReLU activation function, the NNGP kernel is given by

$$k_{\text{nngp}}(\mathbf{x}, \tilde{\mathbf{x}}) := k_{\text{nngp}}^{(L)}(\mathbf{x}, \tilde{\mathbf{x}}) ,$$

where we roughly follow Lee et al. (2019) and recursively define

$$\begin{aligned} k_{\text{nngp}}^{(1)}(\mathbf{x}, \tilde{\mathbf{x}}) &:= \frac{\sigma_w^2}{d} \langle \mathbf{x}, \tilde{\mathbf{x}} \rangle \\ k_{\text{nngp}}^{(l+1)}(\mathbf{x}, \tilde{\mathbf{x}}) &:= \sigma_w^2 f(k_{\text{nngp}}^{(l)}(\mathbf{x}, \mathbf{x}), k_{\text{nngp}}^{(l)}(\mathbf{x}, \tilde{\mathbf{x}}), k_{\text{nngp}}^{(l)}(\tilde{\mathbf{x}}, \tilde{\mathbf{x}})) \\ f(a, b, c) &:= \frac{\sqrt{ac}}{2\pi} \left( \sqrt{1 - u^2} + u(\pi - \arccos(u)) \right) \text{ with } u := \frac{b}{\sqrt{ac}} . \end{aligned}$$

Note that we do not include the  $\sigma_b$  terms here since we initialize the biases to zero unlike Lee et al. (2019).

## Appendix C. Details on Kernel Transformations

In the following, we will discuss additional aspects of various kernel transformations.

### C.1 Gaussian Process Posterior Transformation

Khan et al. (2019) showed that certain posterior approximation methods for NNs turn them into Gaussian processes with the finite-width NTK  $k_{\text{grad}}$ . In the following, we will present a self-contained derivation of this relationship in our framework, including the last-layer kernel  $k_{\parallel}$ . In our exposition, we roughly follow Daxberger et al. (2021). For a Bayesian NN, we impose a prior  $p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta} \mid \mathbf{0}, \lambda^2 \mathbf{I})$  on the NN parameters  $\boldsymbol{\theta}$ . Using an observation model  $y_i = f_{\boldsymbol{\theta}}(\mathbf{x}_i) + \varepsilon_i, \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$  i.i.d., the negative log-likelihood of the data is given by

$$-\log p(\mathcal{Y}_{\text{train}} \mid \mathcal{X}_{\text{train}}, \boldsymbol{\theta}) = C + \frac{1}{2\sigma^2} \sum_{i=1}^{N_{\text{train}}} (y_i - f_{\boldsymbol{\theta}}(\mathbf{x}_i))^2 = C_1 + \frac{1}{2\sigma^2} N_{\text{train}} \mathcal{L}(\boldsymbol{\theta})$$

for some constant  $C_1 \in \mathbb{R}$ . The negative log-posterior is hence given by

$$\tilde{\mathcal{L}}(\boldsymbol{\theta}) := -\log p(\boldsymbol{\theta} \mid \mathcal{Y}_{\text{train}}, \mathcal{X}_{\text{train}})$$

$$\begin{aligned}
 &= \log(Z) - \log p(\mathcal{Y}_{\text{train}} \mid \mathcal{X}_{\text{train}}, \boldsymbol{\theta}) - \log p(\boldsymbol{\theta}) \\
 &= C_2 + \frac{1}{2\sigma^2} N_{\text{train}} \mathcal{L}(\boldsymbol{\theta}) + \frac{1}{2\lambda^2} \|\boldsymbol{\theta}\|_2^2,
 \end{aligned}$$

If  $\boldsymbol{\theta}^*$  minimizes  $\tilde{\mathcal{L}}$ , that is, if  $\boldsymbol{\theta}^*$  is a maximum *a posteriori* (MAP) estimate, we obtain the second-order Taylor approximation

$$\tilde{\mathcal{L}}(\boldsymbol{\theta}) \approx \tilde{\mathcal{L}}(\boldsymbol{\theta}^*) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \mathbf{H}(\boldsymbol{\theta} - \boldsymbol{\theta}^*), \quad \mathbf{H} := \nabla_{\boldsymbol{\theta}}^2 \tilde{\mathcal{L}}(\boldsymbol{\theta}^*),$$

which yields a Gaussian approximation to the posterior, the so-called Laplace approximation (Laplace, 1774; MacKay, 1992a):

$$p(\boldsymbol{\theta} \mid \mathcal{Y}_{\text{train}}, \mathcal{X}_{\text{train}}) \approx \mathcal{N}(\boldsymbol{\theta} \mid \boldsymbol{\theta}^*, \mathbf{H}^{-1}).$$

The Hessian is given by

$$\begin{aligned}
 \mathbf{H} &= \frac{1}{\lambda^2} \mathbf{I} + \frac{1}{2\sigma^2} \sum_{(\mathbf{x}, y) \in \mathcal{D}_{\text{train}}} \nabla_{\boldsymbol{\theta}}^2 (y - f_{\boldsymbol{\theta}^*}(\mathbf{x}))^2 \\
 &= \lambda^{-2} \mathbf{I} + \sigma^{-2} \sum_{(\mathbf{x}, y) \in \mathcal{D}_{\text{train}}} \left( (\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}^*}(\mathbf{x})) (\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}^*}(\mathbf{x}))^\top + (f_{\boldsymbol{\theta}^*}(\mathbf{x}) - y) \nabla_{\boldsymbol{\theta}}^2 f_{\boldsymbol{\theta}^*}(\mathbf{x}) \right).
 \end{aligned}$$

By ignoring the terms  $(f_{\boldsymbol{\theta}^*}(\mathbf{x}) - y) \nabla_{\boldsymbol{\theta}}^2 f_{\boldsymbol{\theta}^*}(\mathbf{x})$ , which are expected to be small since the first factor is small, we arrive at the generalized Gauss-Newton (GGN) approximation to the Hessian (Schraudolph, 2002):

$$\mathbf{H}_{\text{GGN}} = \lambda^{-2} \mathbf{I} + \sigma^{-2} \sum_{(\mathbf{x}, y) \in \mathcal{D}_{\text{train}}} (\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}^*}(\mathbf{x})) (\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}^*}(\mathbf{x}))^\top.$$

By pretending that  $\boldsymbol{\theta}^* = \boldsymbol{\theta}_T$ , i.e. that the parameters at the end of training are the minimizer of  $\tilde{\mathcal{L}}$ , we can relate this to  $\phi_{\text{grad}}$ :

$$\begin{aligned}
 \mathbf{H}_{\text{GGN}} &= \lambda^{-2} \mathbf{I} + \sigma^{-2} \sum_{(\mathbf{x}, y) \in \mathcal{D}_{\text{train}}} \phi_{\text{grad}}(\mathbf{x}) \phi_{\text{grad}}(\mathbf{x})^\top \\
 &= \lambda^{-2} \mathbf{I} + \sigma^{-2} \phi_{\text{grad}}(\mathcal{X}_{\text{train}})^\top \phi_{\text{grad}}(\mathcal{X}_{\text{train}}).
 \end{aligned}$$

If we want to compute the predictive distribution of  $f_{\boldsymbol{\theta}}(\mathbf{x})$  for  $\boldsymbol{\theta} \sim p(\boldsymbol{\theta} \mid \mathcal{X}_{\text{train}})$ , we can further use the linearization

$$f_{\boldsymbol{\theta}}(\mathbf{x}) \approx f_{\boldsymbol{\theta}^*}(\mathbf{x}) + \langle \boldsymbol{\theta} - \boldsymbol{\theta}^*, \nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}^*}(\mathbf{x}) \rangle = f_{\boldsymbol{\theta}^*}(\mathbf{x}) + \langle \boldsymbol{\theta} - \boldsymbol{\theta}^*, \phi_{\text{grad}}(\mathbf{x}) \rangle,$$

which, according to Immer et al. (2021), improves the results for the predictive distribution. The predictive distribution can then be approximated as

$$p(\mathcal{Y} \mid \mathcal{X}, \mathcal{D}_{\text{train}}) \approx \mathcal{N}(\mathcal{Y} \mid f_{\boldsymbol{\theta}^*}(\mathcal{X}), \phi_{\text{grad}}(\mathcal{X}) \mathbf{H}_{\text{GGN}}^{-1} \phi_{\text{grad}}(\mathcal{X})^\top),$$

with the covariance matrix

$$\phi_{\text{grad}}(\mathcal{X}) \mathbf{H}_{\text{GGN}}^{-1} \phi_{\text{grad}}(\mathcal{X})^\top$$

$$\begin{aligned}
 &= \phi_{\text{grad}}(\mathcal{X})(\lambda^{-2}\mathbf{I} + \sigma^{-2}\phi_{\text{grad}}(\mathcal{X}_{\text{train}})^\top \phi_{\text{grad}}(\mathcal{X}_{\text{train}}))^{-1}\phi_{\text{grad}}(\mathcal{X})^\top \\
 &= \sigma^2\lambda\phi_{\text{grad}}(\mathcal{X})(\lambda\phi_{\text{grad}}(\mathcal{X}_{\text{train}})^\top \lambda\phi_{\text{grad}}(\mathcal{X}_{\text{train}}) + \sigma^2\mathbf{I})^{-1}\lambda\phi_{\text{grad}}(\mathcal{X})^\top \\
 &= (\lambda^2k_{\text{grad}})_{\rightarrow\text{post}(\mathcal{X}_{\text{train}},\sigma^2)}(\mathcal{X}, \mathcal{X}) .
 \end{aligned}$$

This demonstrates that  $(\lambda^2k_{\text{grad}})_{\rightarrow\text{post}(\mathcal{X}_{\text{train}},\sigma^2)}(\mathcal{X}, \mathcal{X})$  yields the posterior predictive covariance on  $\mathcal{X}$  of a Bayesian NN under the following approximations:

- (1) The parameter posterior is approximated using the Laplace approximation,
- (2) The Hessian matrix in the Laplace approximation is approximated using the GGN approximation,
- (3) The predictive posterior is further approximated using a linearization of the NN, and
- (4) The MAP estimate is approximated by the trained parameters of the NN.

If we perform Bayesian inference only over the last-layer parameters (i.e.,  $\boldsymbol{\theta} = \tilde{\mathbf{W}}^{(L)}$ ), the derivation above shows instead that the posterior predictive covariance is approximated by  $(\lambda^2k_{\text{ll}})_{\rightarrow\text{post}(\mathcal{X}_{\text{train}},\sigma^2)}(\mathcal{X}, \mathcal{X})$ . Moreover, for the last-layer parameters, the approximations (1) – (3) are exact, since the NN is affine linear in the last-layer parameters, and the approximation (4) does not change the resulting kernel. Such last-layer Bayesian models have been employed, for example, by Lázaro-Gredilla and Figueiras-Vidal (2010); Snoek et al. (2015); Ober and Rasmussen (2019); Kristiadi et al. (2020), and are also known as neural linear models (Ober and Rasmussen, 2019).

Eschenhagen et al. (2021) experimentally demonstrated that taking a mixture of multiple Laplace approximations around different local minima of the loss function can improve uncertainty predictions for Bayesian NNs. If we consider  $N_{\text{ens}}$  local minima  $\boldsymbol{\theta}^{(i)}$  with corresponding base kernels  $k^{(i)}$  and combine the Laplace approximations with uniform weights, we obtain the posterior distributions

$$p(\mathcal{Y} | \mathcal{X}, \mathcal{D}_{\text{train}}) \approx \frac{1}{N_{\text{ens}}} \sum_{i=1}^{N_{\text{ens}}} \mathcal{N}(\mathcal{Y} | f_{\boldsymbol{\theta}^{(i)}}(\mathcal{X}), (\lambda k^{(i)})_{\rightarrow\text{post}(\mathcal{X}_{\text{train}},\sigma^2)}(\mathcal{X}, \mathcal{X})) .$$

By the law of total covariance, we have

$$\begin{aligned}
 \text{Cov}(y_1, y_2 | \mathbf{x}_1, \mathbf{x}_2, \mathcal{D}_{\text{train}}) &\approx \mathbb{E}_{i \sim \mathcal{U}\{1, \dots, N_{\text{ens}}\}} [(\lambda^2 k^{(i)})_{\rightarrow\text{post}(\mathcal{X}_{\text{train}},\sigma^2)}(\mathbf{x}_1, \mathbf{x}_2)] \\
 &\quad + \text{Cov}_{i \sim \mathcal{U}\{1, \dots, N_{\text{ens}}\}}(f_{\boldsymbol{\theta}^{(i)}}(\mathbf{x}_1), f_{\boldsymbol{\theta}^{(i)}}(\mathbf{x}_2)) \\
 &= (\lambda^2 k)_{\rightarrow\text{post}(\mathcal{X}_{\text{train}},\sigma^2) \rightarrow \text{ens}(N_{\text{ens}})}(\mathbf{x}_1, \mathbf{x}_2) \\
 &\quad + \text{Cov}_{i \sim \mathcal{U}\{1, \dots, N_{\text{ens}}\}}(f_{\boldsymbol{\theta}^{(i)}}(\mathbf{x}_1), f_{\boldsymbol{\theta}^{(i)}}(\mathbf{x}_2)) .
 \end{aligned}$$

Hence, the predictive covariance for a mixture of Laplace approximations is approximately given by an ensembled posterior kernel plus the covariance of the ensemble predictions. Note that due to the summation, it is important that the (posterior) kernel is scaled correctly. We leave an experimental evaluation of this approach to future work.

## C.2 Sketching

In the following, we prove the variant of the Johnson-Lindenstrauss theorem mentioned in Section 4.2.3.

**Theorem 1** (Variant of the Johnson-Lindenstrauss Lemma). *Let  $\varepsilon, \delta \in (0, 1)$  and let  $\mathcal{X} \subseteq \mathbb{R}^d$  be finite. If*

$$p \geq 8 \log(|\mathcal{X}|^2 / \delta) / \varepsilon^2, \quad (14)$$

*then the following bound on all pairwise distances holds with probability  $\geq 1 - \delta$  for the Gaussian sketch in Eq. (12):*

$$\forall \mathbf{x}, \tilde{\mathbf{x}} \in \mathcal{X} : (1 - \varepsilon)d_k(\mathbf{x}, \tilde{\mathbf{x}}) \leq d_{k \rightarrow \text{sketch}(p)}(\mathbf{x}, \tilde{\mathbf{x}}) \leq (1 + \varepsilon)d_k(\mathbf{x}, \tilde{\mathbf{x}}). \quad (15)$$

**Proof** By Theorem 1 from Arriaga and Vempala (1999), the following bound holds for fixed  $\mathbf{x}, \tilde{\mathbf{x}} \in \mathcal{X}$  with probability<sup>6</sup>  $\geq 1 - 2e^{-\varepsilon^2 p/8}$ :

$$(1 - \varepsilon)d_k(\mathbf{x}, \tilde{\mathbf{x}})^2 \leq d_{k \rightarrow \text{sketch}(p)}(\mathbf{x}, \tilde{\mathbf{x}})^2 \leq (1 + \varepsilon)d_k(\mathbf{x}, \tilde{\mathbf{x}})^2. \quad (26)$$

Now, Eq. (26) implies Eq. (15) for a single pair  $\mathbf{x}, \tilde{\mathbf{x}}$  because of

$$(1 - \varepsilon)^2 \leq (1 - \varepsilon) \leq (1 + \varepsilon) \leq (1 + \varepsilon)^2.$$

To obtain the bound for all pairs  $(\mathbf{x}, \tilde{\mathbf{x}})$ , we note that the bound is trivial for pairs  $(\mathbf{x}, \mathbf{x})$  and it is equivalent for the pairs  $(\mathbf{x}, \tilde{\mathbf{x}})$  and  $(\tilde{\mathbf{x}}, \mathbf{x})$ . Hence, we only have to consider less than  $\frac{|\mathcal{X}|^2}{2}$  pairs. By the union bound, the probability that Eq. (15) holds for all pairs  $(\mathbf{x}, \tilde{\mathbf{x}})$  is at least

$$1 - 2 \frac{|\mathcal{X}|^2}{2} e^{-\varepsilon^2 p/8} \stackrel{\text{Eq. (14)}}{\geq} 1 - \delta. \quad \blacksquare$$

### C.3 ACS Random Features Transformation

We will now derive the form of  $k_{\rightarrow \text{acs}}$  presented in Section 4.2.5. Partially following the notation of Pinsler et al. (2019), we want to compute

$$k_{\rightarrow \text{acs}}(\mathbf{x}_n, \mathbf{x}_m) := \langle \mathcal{L}_n, \mathcal{L}_m \rangle = \mathbb{E}_{\boldsymbol{\theta} \sim p(\boldsymbol{\theta} | \mathcal{D}_{\text{train}})}[\mathcal{L}_n(\boldsymbol{\theta}) \mathcal{L}_m(\boldsymbol{\theta})],$$

where

$$\mathcal{L}_m(\boldsymbol{\theta}) = \mathbb{E}_{y_m \sim p(\cdot | \mathbf{x}_m, \mathcal{D}_{\text{train}})}[\log p(y_m | \mathbf{x}_m, \boldsymbol{\theta})] + \mathbb{H}[y_m | \mathbf{x}_m, \mathcal{D}_{\text{train}}]$$

with  $\mathbb{H}[y_m | \mathbf{x}_m, \mathcal{D}_{\text{train}}]$  denoting the conditional entropy of  $y_m$  given  $\mathbf{x}_m$  and  $\mathcal{D}_{\text{train}}$ .

In a GP model without hyper-prior on  $\sigma^2$ , we have

$$\begin{aligned} p(y_m | \mathbf{x}_m, \boldsymbol{\theta}) &= \mathcal{N}(y_m | f_{\boldsymbol{\theta}_T}(\mathbf{x}_m) + \boldsymbol{\theta}^\top \phi_{\rightarrow \text{scale}(\mathcal{X}_{\text{train}})}(\mathbf{x}_m), \sigma^2) \\ p(y_m | \mathbf{x}_m, \mathcal{D}_{\text{train}}) &= \mathcal{N}(y_m | f_{\boldsymbol{\theta}_T}(\mathbf{x}_m), k_{\rightarrow \mathcal{X}_{\text{train}}}(\mathbf{x}_m, \mathbf{x}_m) + \sigma^2) \\ \mathbb{H}(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})) &= \frac{n}{2} \ln(2\pi e) + \frac{1}{2} \ln(\det(\boldsymbol{\Sigma})) \quad \text{for } \boldsymbol{\mu} \in \mathbb{R}^n, \boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}, \end{aligned}$$

which allows us to derive

$$\mathbb{H}[y_m | \mathbf{x}_m, \mathcal{D}_{\text{train}}] = \frac{1}{2} + \frac{1}{2} \log(2\pi(k_{\rightarrow \mathcal{X}_{\text{train}}}(\mathbf{x}_m, \mathbf{x}_m) + \sigma^2)).$$

---

6. We inserted the factor 2 in front of  $e^{-\varepsilon^2 p/8}$  that has been forgotten in their Theorem 1.



By shifting the prior and the log-likelihood by  $f_{\theta_T}(\mathbf{x}_m)$ , we obtain

$$\begin{aligned}
 & \mathbb{E}_{y_m \sim p(\cdot | \mathbf{x}_m, \mathcal{D}_0)}[\log p(y_m | \mathbf{x}_m, \boldsymbol{\theta})] \\
 &= \mathbb{E}_{y_m \sim \mathcal{N}(0, k_{\rightarrow \mathcal{X}_{\text{train}}}(\mathbf{x}_m, \mathbf{x}_m) + \sigma^2)}[\log \mathcal{N}(y_m | \boldsymbol{\theta}^\top \phi_{\rightarrow \text{scale}(\mathcal{X}_{\text{train}})}(\mathbf{x}_m), \sigma^2)] \\
 &= -\frac{1}{2} \log(2\pi\sigma^2) - \mathbb{E}_{y_m \sim \mathcal{N}(0, k_{\rightarrow \mathcal{X}_{\text{train}}}(\mathbf{x}_m, \mathbf{x}_m) + \sigma^2)} \frac{(y_m - \boldsymbol{\theta}^\top \phi_{\rightarrow \text{scale}(\mathcal{X}_{\text{train}})}(\mathbf{x}_m))^2}{2\sigma^2} \\
 &= -\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \left( (\boldsymbol{\theta}^\top \phi_{\rightarrow \text{scale}(\mathcal{X}_{\text{train}})}(\mathbf{x}_m))^2 + k_{\rightarrow \mathcal{X}_{\text{train}}}(\mathbf{x}_m, \mathbf{x}_m) + \sigma^2 \right).
 \end{aligned}$$

Therefore, we have  $k_{\rightarrow \text{acs}}(\mathbf{x}, \tilde{\mathbf{x}}) := \mathbb{E}_{\boldsymbol{\theta} \sim P(\boldsymbol{\theta} | \mathcal{D}_{\text{train}})}[f_{\text{acs}}(\mathbf{x}, \boldsymbol{\theta}) f_{\text{acs}}(\tilde{\mathbf{x}}, \boldsymbol{\theta})]$  with

$$\begin{aligned}
 f_{\text{acs}}(\mathbf{x}_m, \boldsymbol{\theta}) := \mathcal{L}_m(\boldsymbol{\theta}) &= \frac{1}{2} \log \left( 1 + \frac{k_{\rightarrow \mathcal{X}_{\text{train}}}(\mathbf{x}_m, \mathbf{x}_m)}{\sigma^2} \right) \\
 &\quad - \frac{(\boldsymbol{\theta}^\top \phi_{\rightarrow \text{scale}(\mathcal{X}_{\text{train}})}(\mathbf{x}_m))^2 + k_{\rightarrow \mathcal{X}_{\text{train}}}(\mathbf{x}_m, \mathbf{x}_m)}{2\sigma^2}.
 \end{aligned}$$

Moreover, it follows from Eq. (4) in Pinsler et al. (2019) that

$$f_{\text{pool}}(\boldsymbol{\theta}) - f_{\text{batch}}(\boldsymbol{\theta}) = \sum_{\mathbf{x}_m \in \mathcal{X}_{\text{pool}} \setminus \mathcal{X}_{\text{batch}}} \mathcal{L}_m(\boldsymbol{\theta}) = \sum_{\mathbf{x} \in \mathcal{X}_{\text{pool}} \setminus \mathcal{X}_{\text{batch}}} f_{\text{acs}}(\mathbf{x}, \boldsymbol{\theta}).$$

## Appendix D. Details on Selection Methods

In this section, we will provide efficiency-focused pseudocode for all selection methods and analyze its runtime and memory complexity. Hereby, we will neglect that the required integer bit size for indexing elements of  $\mathcal{X}_{\text{pool}}$  and  $\mathcal{X}_{\text{train}}$  grows logarithmically with  $N_{\text{pool}} + N_{\text{train}}$ . For some selection methods, we will additionally discuss relations to the literature and theoretical properties. Section D.1 will first provide a pseudocode structure for iterative selection, where the missing components are then specified for the respective selection methods in the subsequent sections. The following notation will be used throughout this section:

We allow having vectors and matrices indexed by points  $\mathbf{x}$  instead of indices  $i \in \mathbb{N}$ , which we write with square brackets as  $\mathbf{v}[\mathbf{x}]$  or  $\mathbf{M}[\mathbf{x}, \tilde{\mathbf{x}}]$ . In a practical implementation, where the points  $\mathbf{x} \in \mathcal{X}_{\text{pool}}$  are for example numbered as  $\mathbf{x}_1, \dots, \mathbf{x}_{N_{\text{pool}}}$ , one may simply use  $\mathbf{v}[i]$  instead of  $\mathbf{v}[\mathbf{x}_i]$ . Again, we assume in pseudocode that all  $\mathbf{x}$  are distinct, such that we can use set notation, but identical copies of  $\mathbf{x}$  should be treated as distinct. This problem also disappears when using indices. We denote by  $\mathbf{u} \odot \mathbf{v}$  the element-wise (Hadamard) product of the vectors  $\mathbf{u}$  and  $\mathbf{v}$ . Whenever an argmax is not unique, we leave the choice of the maximizer to the implementation.

### D.1 Iterative Selection Scheme

While our iterative selection template shown in Algorithm 3 is sufficient for a high-level understanding, it is not well-suited for an efficient implementation. To this end, we present a more detailed iterative selection template in Algorithm 3, which closely matches our open-source implementation. The template in Algorithm 3 involves three methods called INIT, ADD, and NEXT, which are allowed to have side effects, i.e. access and modify common variables. For each selection method except RANDOM and MAXDIAG, our implementation

---

**Algorithm D.1** Iterative selection algorithm template involving three customizable functions INIT, ADD and NEXT that are allowed to have side effects (i.e., read/write variables in SELECT).

---

```

function SELECT( $k$ ,  $\mathcal{X}_{\text{train}}$ ,  $\mathcal{X}_{\text{pool}}$ ,  $N_{\text{batch}}$ , mode  $\in \{\text{P}, \text{TP}\}$ )
     $\mathcal{X}_{\text{mode}} \leftarrow \mathcal{X}_{\text{train}}$  if mode = TP else  $\emptyset$ 
     $\mathcal{X}_{\text{cand}} := \mathcal{X}_{\text{mode}} \cup \mathcal{X}_{\text{pool}}$ 
     $\mathcal{X}_{\text{batch}} \leftarrow \emptyset$ 
    INIT
    for  $\mathbf{x}$  in  $\mathcal{X}_{\text{mode}}$  do
        ADD( $\mathbf{x}$ )
    end for
    for  $i$  from 1 to  $N_{\text{batch}}$  do
         $\mathbf{x} \leftarrow \text{NEXT}$ 
        if  $\mathbf{x} \in \mathcal{X}_{\text{batch}} \cup \mathcal{X}_{\text{train}}$  (failed selection) then
            fill up  $\mathcal{X}_{\text{batch}}$  with  $N_{\text{batch}} - |\mathcal{X}_{\text{batch}}|$  random samples from  $\mathcal{X}_{\text{pool}} \setminus \mathcal{X}_{\text{batch}}$  and
            return  $\mathcal{X}_{\text{batch}}$ 
        end if
         $\mathcal{X}_{\text{batch}} \leftarrow \mathcal{X}_{\text{batch}} \cup \{\mathbf{x}\}$ 
        ADD( $\mathbf{x}$ )
    end for
    return  $\mathcal{X}_{\text{batch}}$ 
end function

```

---

directly mirrors this structure, containing a class providing the three methods together with SELECT from Algorithm 3.

## D.2 RANDOM

We implement RANDOM by taking the first  $N_{\text{batch}}$  indices out of a random permutation of  $\{1, \dots, N_{\text{pool}}\}$ . Since there are  $N_{\text{pool}}!$  possible random permutations, this requires at least  $\log_2(N_{\text{pool}}!) = \mathcal{O}(N_{\text{pool}} \log N_{\text{pool}})$  random bits, so the runtime for this suboptimal implementation is, in theory, “only”  $\mathcal{O}(N_{\text{pool}} \log N_{\text{pool}})$ , which is still extremely fast in practice. The memory complexity for our suboptimal implementation is  $\mathcal{O}(N_{\text{pool}})$ .

## D.3 MAXDIAG

A simple implementation of MAXDIAG is shown in Algorithm D.2. The runtime of this implementation is  $\mathcal{O}(N_{\text{pool}} \log N_{\text{pool}})$  due to sorting. While other algorithms might be faster for  $N_{\text{batch}} \ll N_{\text{pool}}$ , the runtime is already very fast in practice. The memory complexity is  $\mathcal{O}(N_{\text{pool}})$ .

---

**Algorithm D.2** MAXDIAG pseudocode implementation using Algorithm D.1.

---

**function** INIT

Sort elements in  $\mathcal{X}_{\text{pool}}$  as  $\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_{N_{\text{pool}}}$  such that  $k(\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_1) \geq \dots \geq k(\tilde{\mathbf{x}}_{N_{\text{pool}}}, \tilde{\mathbf{x}}_{N_{\text{pool}}})$

**end function**

**function** ADD( $\mathbf{x}$ )

**end function**

**function** NEXT

return  $\tilde{\mathbf{x}}_i$

**end function**

---

## D.4 MAXDET

### D.4.1 EQUIVALENCE OF MAXDET TO NON-BATCH MODE ACTIVE LEARNING WITH FIXED KERNEL

Using the Schur determinant formula

$$\det \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix} = \det(\mathbf{A}) \det(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}) ,$$

we can compute

$$\begin{aligned} & \det(k(\mathcal{X} \cup \{\mathbf{x}\}, \mathcal{X} \cup \{\mathbf{x}\}) + \sigma^2 \mathbf{I}) \\ &= \det \begin{pmatrix} k(\mathcal{X}, \mathcal{X}) + \sigma^2 \mathbf{I} & k(\mathcal{X}, \mathbf{x}) \\ k(\mathbf{x}, \mathcal{X}) & k(\mathbf{x}, \mathbf{x}) + \sigma^2 \end{pmatrix} \\ &= \det(k(\mathcal{X}, \mathcal{X}) + \sigma^2 \mathbf{I}) \det(k(\mathbf{x}, \mathbf{x}) + \sigma^2 - k(\mathbf{x}, \mathcal{X})(k(\mathcal{X}, \mathcal{X}) + \sigma^2 \mathbf{I})^{-1}k(\mathcal{X}, \mathbf{x})) \\ &= \det(k(\mathcal{X}, \mathcal{X}) + \sigma^2 \mathbf{I}) \cdot (\sigma^2 + k_{\rightarrow \text{post}}(\mathcal{X}, \sigma^2)(\mathbf{x}, \mathbf{x})) . \end{aligned} \tag{27}$$

This shows that

$$\operatorname{argmax}_{\mathbf{x} \in \mathcal{X}_{\text{rem}}} k_{\rightarrow \text{post}}(\mathcal{X}_{\text{sel}}, \sigma^2)(\mathbf{x}, \mathbf{x}) = \operatorname{argmax}_{\mathbf{x} \in \mathcal{X}_{\text{rem}}} \det(k(\mathcal{X}_{\text{sel}} \cup \{\mathbf{x}\}, \mathcal{X}_{\text{sel}} \cup \{\mathbf{x}\}) + \sigma^2 \mathbf{I}) .$$

### D.4.2 EQUIVALENCE OF MAXDET TO BATCHBALD ON A GP

As in Section 4.2.2, we consider a GP model in feature space, given by  $y_i = \mathbf{w}^\top \phi(\mathbf{x}_i) + \varepsilon_i$  with weight prior  $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$  and i.i.d. observation noise  $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ . The objective of BatchBALD (Kirsch et al., 2019) is to maximize the mutual information

$$a(\mathcal{X}_{\text{batch}}) := \mathbb{H}(\mathcal{Y}_{\text{batch}} \mid \mathcal{X}_{\text{batch}}, \mathcal{D}_{\text{train}}) - \mathbb{E}_{\mathbf{w} \sim p(\mathbf{w} \mid \mathcal{D}_{\text{train}})} \mathbb{H}(\mathcal{Y}_{\text{batch}} \mid \mathcal{X}_{\text{batch}}, \mathcal{D}_{\text{train}}, \mathbf{w}) ,$$

where  $\mathbb{H}$  refers to the (conditional) entropy. Writing  $\mathcal{Y}_{\text{batch}}^* := \mathbb{E}[\mathcal{Y}_{\text{batch}} \mid \mathcal{X}_{\text{batch}}, \mathcal{D}_{\text{train}}]$ , we have

$$\begin{aligned} p(\mathcal{Y}_{\text{batch}} \mid \mathcal{X}_{\text{batch}}, \mathcal{D}_{\text{train}}) &= \mathcal{N}(\mathcal{Y}_{\text{batch}} \mid \mathcal{Y}_{\text{batch}}^*, k(\mathcal{X}_{\text{batch}}, \mathcal{X}_{\text{batch}}) + \sigma^2 \mathbf{I}) \\ p(\mathcal{Y}_{\text{batch}} \mid \mathcal{X}_{\text{batch}}, \mathcal{D}_{\text{train}}, \mathbf{w}) &= \mathcal{N}(\mathcal{Y}_{\text{batch}} \mid \phi(\mathcal{X}_{\text{train}})^\top \mathbf{w}, \sigma^2 \mathbf{I}) \end{aligned}$$

$$\mathbb{H}(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})) = \frac{n}{2} \ln(2\pi e) + \frac{1}{2} \ln(\det(\boldsymbol{\Sigma})) \quad \text{for } \boldsymbol{\mu} \in \mathbb{R}^n, \boldsymbol{\Sigma} \in \mathbb{R}^{n \times n} .$$

Hence, we can compute

$$a(\mathcal{X}_{\text{batch}}) = \frac{1}{2} \ln(\det(k(\mathcal{X}_{\text{batch}}, \mathcal{X}_{\text{batch}}) + \sigma^2 \mathbf{I})) - N_{\text{batch}} \ln(\sigma) . \quad (28)$$

This shows that greedy maximization of the BatchBALD acquisition function, as proposed by Kirsch et al. (2019), is equivalent to MAXDET. Kirsch et al. (2019) showed that for any Bayesian model, greedy optimization of  $a$  is suboptimal by a factor of at most  $(1 - 1/e)$ . By applying this result to the GP model, we obtain that the same suboptimality bound applies to MAXDET for the form of  $a$  given in Eq. (28).

#### D.4.3 EQUIVALENCE OF MAXDET TO THE P-GREEDY ALGORITHM

In our notation, the P-greedy algorithm (De Marchi et al., 2005) can be written as

$$\text{NEXTSAMPLE}(k, \mathcal{X}_{\text{sel}}, \mathcal{X}_{\text{rem}}) = \underset{\mathbf{x} \in \mathcal{X}_{\text{rem}}}{\operatorname{argmax}} P_{k, \mathcal{X}_{\text{sel}}}(\mathbf{x}) ,$$

where by Lemma 4.1 in De Marchi et al. (2005), the non-negative power function  $P_{k, \mathcal{X}_{\text{sel}}}$  can be written as

$$P_{k, \mathcal{X}_{\text{sel}}}(\mathbf{x})^2 = k(\mathbf{x}, \mathbf{x}) - k(\mathbf{x}, \mathcal{X}_{\text{sel}})k(\mathcal{X}_{\text{sel}}, \mathcal{X}_{\text{sel}})^{-1}k(\mathcal{X}_{\text{sel}}, \mathbf{x}) .$$

A calculation analogous to Eq. (27) therefore shows that P-greedy is equivalent to MAXDET with  $\sigma^2 = 0$ .

#### D.4.4 RELATION TO THE GREEDY ALGORITHM FOR D-OPTIMAL DESIGN

In our notation, the D-optimal design problem (Wald, 1943) is to maximize the determinant  $\det(\phi(\mathcal{X}_{\text{sel}})^\top \phi(\mathcal{X}_{\text{sel}}))$ , which can only be nonzero in the case  $N_{\text{sel}} \geq d_{\text{feat}}$ . It can be seen as the  $\sigma \rightarrow 0$  limit for  $N_{\text{sel}} \geq d_{\text{feat}}$  of the determinant-maximization objective that motivates MAXDET since an eigenvalue-based argument shows that

$$\det(\phi(\mathcal{X}_{\text{sel}})^\top \phi(\mathcal{X}_{\text{sel}}) + \sigma^2 \mathbf{I}) = \sigma^{d_{\text{feat}} - N_{\text{sel}}} \det(k(\mathcal{X}_{\text{sel}}, \mathcal{X}_{\text{sel}}) + \sigma^2 \mathbf{I}) .$$

In this sense, the corresponding greedy algorithm (Wynn, 1970) is the underparameterized ( $d_{\text{feat}} \leq N_{\text{sel}}$ ) analog to the P-greedy algorithm, since the latter can only be well-defined in the overparameterized regime ( $d_{\text{feat}} \geq N_{\text{sel}}$ ). Some guarantees for this greedy algorithm are given by Wynn (1970) and Madan et al. (2019). While the classical D-optimal design uses  $\sigma = 0$ , Bayesian D-optimal design uses  $\sigma > 0$  and is thus even more directly related to MAXDET (Chaloner and Verdinelli, 1995).

#### D.4.5 KERNEL-SPACE IMPLEMENTATION OF MAXDET

In the following, we want to derive an efficient kernel-space implementation of MAXDET. Let  $\mathcal{X}_{\text{cand}} := \mathcal{X}_{\text{mode}} \cup \mathcal{X}_{\text{pool}}$ . We perform a partial pivoted Cholesky decomposition of the matrix  $\mathbf{M} = k(\mathcal{X}_{\text{cand}}, \mathcal{X}_{\text{cand}}) + \sigma^2 \mathbf{I}$ , which has been suggested for P-greedy by Pazouki and

Schaback (2011) and in the context of determinantal point processes by Chen et al. (2018). We denote submatrices of  $\mathbf{M}$  for example by  $\mathbf{M}[\mathcal{X}, \mathbf{x}] := (M_{\tilde{\mathbf{x}}, \mathbf{x}})_{\tilde{\mathbf{x}} \in \mathcal{X}}$ .

Suppose that at the current step, the points  $\mathcal{X} := \mathcal{X}_{\text{mode}} \cup \mathcal{X}_{\text{batch}}$  have already been added. Consider the Cholesky decomposition  $\mathbf{M}[\mathcal{X}, \mathcal{X}] = \mathbf{L}(\mathcal{X})\mathbf{L}(\mathcal{X})^\top$ . Then, the Cholesky decomposition for  $\mathcal{X} \cup \{\mathbf{x}\}$  is of the form

$$\begin{aligned} \begin{pmatrix} \mathbf{M}[\mathcal{X}, \mathcal{X}] & \mathbf{M}[\mathcal{X}, \mathbf{x}] \\ \mathbf{M}[\mathbf{x}, \mathcal{X}] & \mathbf{M}[\mathbf{x}, \mathbf{x}] \end{pmatrix} &= \mathbf{L}(\mathcal{X} \cup \{\mathbf{x}\})\mathbf{L}(\mathcal{X} \cup \{\mathbf{x}\})^\top \\ &= \begin{pmatrix} \mathbf{L}(\mathcal{X}) & \mathbf{0} \\ \mathbf{b}(\mathcal{X}, \mathbf{x})^\top & \sqrt{c(\mathcal{X}, \mathbf{x})} \end{pmatrix} \begin{pmatrix} \mathbf{L}(\mathcal{X}) & \mathbf{0} \\ \mathbf{b}(\mathcal{X}, \mathbf{x})^\top & \sqrt{c(\mathcal{X}, \mathbf{x})} \end{pmatrix}^\top, \end{aligned} \quad (29)$$

which implies  $\mathbf{b}(\mathcal{X}, \mathbf{x}) = \mathbf{L}(\mathcal{X})^{-1}\mathbf{M}[\mathcal{X}, \mathbf{x}]$  and  $c(\mathcal{X}, \mathbf{x}) = \mathbf{M}[\mathbf{x}, \mathbf{x}] - \|\mathbf{b}(\mathcal{X}, \mathbf{x})\|_2^2$ . Using the general inversion formula for block-triangular matrices given by

$$\begin{pmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{B} & \mathbf{C} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{A}^{-1} & \mathbf{0} \\ -\mathbf{C}^{-1}\mathbf{B}\mathbf{A}^{-1} & \mathbf{C}^{-1} \end{pmatrix},$$

we obtain

$$\begin{aligned} \mathbf{b}(\mathcal{X} \cup \{\mathbf{x}\}, \tilde{\mathbf{x}}) &= \mathbf{L}(\mathcal{X} \cup \{\mathbf{x}\})^{-1}\mathbf{M}[\mathcal{X} \cup \{\mathbf{x}\}, \tilde{\mathbf{x}}] \\ &= \begin{pmatrix} \mathbf{L}(\mathcal{X})^{-1} & \mathbf{0} \\ -c(\mathcal{X}, \mathbf{x})^{-1/2}\mathbf{b}(\mathcal{X}, \mathbf{x})^\top \mathbf{L}(\mathcal{X})^{-1} & c(\mathcal{X}, \mathbf{x})^{-1/2} \end{pmatrix} \begin{pmatrix} \mathbf{M}[\mathcal{X}, \tilde{\mathbf{x}}] \\ \mathbf{M}[\mathbf{x}, \tilde{\mathbf{x}}] \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{b}(\mathcal{X}, \tilde{\mathbf{x}}) \\ c(\mathcal{X}, \mathbf{x})^{-1/2}(\mathbf{M}[\mathbf{x}, \tilde{\mathbf{x}}] - \mathbf{b}(\mathcal{X}, \mathbf{x})^\top \mathbf{b}(\mathcal{X}, \tilde{\mathbf{x}})) \end{pmatrix} \end{aligned} \quad (30)$$

and therefore

$$c(\mathcal{X} \cup \{\mathbf{x}\}, \tilde{\mathbf{x}}) = c(\mathcal{X}, \tilde{\mathbf{x}}) - c(\mathcal{X}, \mathbf{x})^{-1}(\mathbf{M}[\mathbf{x}, \tilde{\mathbf{x}}] - \mathbf{b}(\mathcal{X}, \mathbf{x})^\top \mathbf{b}(\mathcal{X}, \tilde{\mathbf{x}}))^2. \quad (31)$$

With the above considerations, we can implement MAXDET as follows, which is given as pseudocode in Algorithm D.3:

- For INIT, we initialize  $\mathbf{b}(\emptyset, \mathbf{x})$  to an empty vector and  $c(\emptyset, \mathbf{x}) = M[\mathbf{x}, \mathbf{x}]$ . We do not precompute  $\mathbf{M}$  since not all entries of  $\mathbf{M}$  will be used, otherwise, the runtime complexity would be quadratic in  $N_{\text{cand}}$ .
- For NEXT, note that by Eq. (29),

$$\det \mathbf{M}[\mathcal{X} \cup \{\mathbf{x}\}, \mathcal{X} \cup \{\mathbf{x}\}] = c(\mathcal{X}, \mathbf{x}) \cdot \det(\mathbf{L}(\mathcal{X}))^2,$$

hence

$$\operatorname{argmax}_{\mathbf{x} \in \mathcal{X}_{\text{pool}} \setminus \mathcal{X}_{\text{batch}}} \det \mathbf{M}[\mathcal{X} \cup \{\mathbf{x}\}, \mathcal{X} \cup \{\mathbf{x}\}] = \operatorname{argmax}_{\mathbf{x} \in \mathcal{X}_{\text{pool}} \setminus \mathcal{X}_{\text{batch}}} c(\mathcal{X}, \mathbf{x}).$$

- For ADD, we update the  $\mathbf{b}$  and  $c$  values as in Eq. (30) and Eq. (31).

Regarding the runtime complexity, the ADD step is clearly the most expensive one, requiring  $\mathcal{O}(N_{\text{cand}}N_{\text{sel}})$  operations for computing  $\mathbf{B}^\top \mathbf{B}[\cdot, \mathbf{x}]$  and  $\mathcal{O}(N_{\text{cand}}T_k)$  operations for computing  $k(\mathcal{X}_{\text{cand}}, \mathbf{x})$ . Since ADD is called  $N_{\text{sel}}$  times, the total runtime complexity is therefore  $\mathcal{O}(N_{\text{cand}}N_{\text{sel}}(T_k + N_{\text{sel}}))$ . The total memory complexity is  $\mathcal{O}(N_{\text{cand}}N_{\text{sel}})$ , required for storing  $\mathbf{B}$ .

---

**Algorithm D.3** MAXDET pseudocode implementation in kernel space for given  $\sigma^2 \geq 0$  using Algorithm D.1.

---

**function** INIT

$\mathbf{B} \leftarrow$  empty  $0 \times \mathcal{X}_{\text{cand}}$  matrix

$\mathbf{c} \leftarrow (k(\mathbf{x}, \mathbf{x}) + \sigma^2)_{\mathbf{x} \in \mathcal{X}_{\text{cand}}}$

**end function**

**function** ADD( $\mathbf{x}$ )

$\mathbf{v} \leftarrow \sqrt{\mathbf{c}} \odot ((k(\mathcal{X}_{\text{cand}}, \mathbf{x}) + \sigma^2 \mathbf{I}[\mathcal{X}_{\text{cand}}, \mathbf{x}] - \mathbf{B}^\top \mathbf{B}[\cdot, \mathbf{x}])) \quad \triangleright \sqrt{\mathbf{c}}$  should be understood element-wise

$\mathbf{B} \leftarrow \begin{pmatrix} \mathbf{B} \\ \mathbf{v}^\top \end{pmatrix}$

$\mathbf{c} \leftarrow \mathbf{c} - \mathbf{v} \odot \mathbf{v}$

**end function**

**function** NEXT

**return**  $\operatorname{argmax}_{\mathbf{x} \in \mathcal{X}_{\text{pool}} \setminus \mathcal{X}_{\text{batch}}} \mathbf{c}[\mathbf{x}]$

**end function**

---

#### D.4.6 FEATURE-SPACE IMPLEMENTATION OF MAXDET

In cases where the feature space dimension  $d_{\text{feat}}$  of  $\phi$  is smaller than the number  $N_{\text{sel}}$  of added samples, it can be beneficial to implement MAXDET in feature space instead. For  $\mathcal{X}_{\text{sel}} := \mathcal{X}_{\text{mode}} \cup \mathcal{X}_{\text{batch}}$ , we want to compute

$$\operatorname{argmax}_{\mathbf{x} \in \mathcal{X}_{\text{pool}} \setminus \mathcal{X}_{\text{batch}}} k_{\rightarrow \text{post}}(\mathcal{X}_{\text{sel}}, \sigma^2)(\mathbf{x}, \mathbf{x}) .$$

Now, from Eq. (9), we know that

$$k_{\rightarrow \text{post}}(\mathcal{X}_{\text{sel}}, \sigma^2)(\mathbf{x}, \mathbf{x}) = \phi(\mathbf{x})^\top \hat{\Sigma}_{\mathcal{X}_{\text{sel}}}^{-1} \phi(\mathbf{x}), \quad \hat{\Sigma}_{\mathcal{X}_{\text{sel}}} := \sigma^{-2} \phi(\mathcal{X}_{\text{sel}})^\top \phi(\mathcal{X}_{\text{sel}}) + \mathbf{I} .$$

Adding a point to  $\mathcal{X}_{\text{sel}}$  leads to a rank-1 update of  $\hat{\Sigma}_{\mathcal{X}_{\text{sel}}}$ , and the corresponding update of  $\hat{\Sigma}_{\mathcal{X}_{\text{sel}}}^{-1}$  can be computed using the Sherman-Morrison formula. This gives rise to three approaches towards implementing MAXDET in feature space:

- (1) Keep track of  $\phi(\mathbf{x})$  and  $\hat{\Sigma}_{\mathcal{X}_{\text{sel}}}^{-1}$ . Compute  $k_{\rightarrow \text{post}}(\mathcal{X}_{\text{sel}}, \sigma^2)(\mathbf{x}, \mathbf{x}) = \phi(\mathbf{x})^\top \hat{\Sigma}_{\mathcal{X}_{\text{sel}}}^{-1} \phi(\mathbf{x})$  in each step.
- (2) Keep track of  $\phi(\mathbf{x})$  and  $\psi(\mathbf{x}) := \hat{\Sigma}_{\mathcal{X}_{\text{sel}}}^{-1} \phi(\mathbf{x})$ . Compute

$$k_{\rightarrow \text{post}}(\mathcal{X}_{\text{sel}}, \sigma^2)(\mathbf{x}, \mathbf{x}) = \phi(\mathbf{x})^\top \psi(\mathbf{x})$$

in each step.

- (3) Keep track of  $\phi_{\rightarrow \text{post}}(\mathcal{X}_{\text{sel}}, \sigma^2)(\mathbf{x})$ , one possible realization of which is  $\phi_{\rightarrow \text{post}}(\mathcal{X}_{\text{sel}}, \sigma^2)(\mathbf{x}) = \hat{\Sigma}_{\mathcal{X}_{\text{sel}}}^{-1/2} \phi(\mathbf{x})$ . Compute  $k_{\rightarrow \text{post}}(\mathcal{X}_{\text{sel}}, \sigma^2)(\mathbf{x}, \mathbf{x}) = \phi_{\rightarrow \text{post}}(\mathcal{X}_{\text{sel}}, \sigma^2)(\mathbf{x})^\top \phi_{\rightarrow \text{post}}(\mathcal{X}_{\text{sel}}, \sigma^2)(\mathbf{x})$  in each step.

Option (1) is less computationally efficient than (2), (3) since one needs to compute matrix-vector products instead of only inner products. Version (2) and (3) are similar, but here we favor (3) since it only requires storing one vector instead of two for each  $\mathbf{x}$ . Since

$$\phi_{\rightarrow\text{post}(\mathcal{X}_{\text{sel}} \cup \{\mathbf{x}\}, \sigma^2)} = \phi_{\rightarrow\text{post}(\mathcal{X}_{\text{sel}}, \sigma^2) \rightarrow \text{post}(\{\mathbf{x}\}, \sigma^2)},$$

we will now consider how to efficiently compute a single posterior update  $\phi_{\rightarrow\text{post}(\{\mathbf{x}\}, \sigma^2)}$ . To this end, we first consider how to compute matrix square roots of specific rank-1 updates:

**Lemma D.1.** *Let  $\mathbf{v} \in \mathbb{R}^p$  and let  $c \geq -\frac{1}{\mathbf{v}^\top \mathbf{v}}$ . Then,*

$$\mathbf{I} + c\mathbf{v}\mathbf{v}^\top = \left( \mathbf{I} + \frac{c}{1 + \sqrt{1 + c\mathbf{v}^\top \mathbf{v}}} \mathbf{v}\mathbf{v}^\top \right)^2.$$

**Proof** Due to the condition on  $c$ , the square root is well-defined. We have

$$\left( \mathbf{I} + \frac{c}{1 + \sqrt{1 + c\mathbf{v}^\top \mathbf{v}}} \mathbf{v}\mathbf{v}^\top \right)^2 = \mathbf{I} + C\mathbf{v}\mathbf{v}^\top,$$

where

$$C = \frac{2c}{1 + \sqrt{1 + c\mathbf{v}^\top \mathbf{v}}} + \frac{c^2 \mathbf{v}^\top \mathbf{v}}{(1 + \sqrt{1 + c\mathbf{v}^\top \mathbf{v}})^2} = \frac{2c(1 + \sqrt{1 + c\mathbf{v}^\top \mathbf{v}}) + c^2 \mathbf{v}^\top \mathbf{v}}{2 + 2\sqrt{1 + c\mathbf{v}^\top \mathbf{v}} + c\mathbf{v}^\top \mathbf{v}} = c. \quad \blacksquare$$

The following proposition shows how to update the posterior feature map after observing a point  $\mathbf{x}$ :

**Proposition D.2** (Forward update). *Let  $\sigma^2 > 0$ , let  $k$  be a kernel and let  $\tilde{k} := k_{\rightarrow\text{post}(\{\mathbf{x}\}, \sigma^2)}$ . Then,*

$$\tilde{k}(\mathbf{x}', \mathbf{x}'') = k(\mathbf{x}', \mathbf{x}'') - k(\mathbf{x}', \mathbf{x})(k(\mathbf{x}, \mathbf{x}) + \sigma^2 \mathbf{I})^{-1} k(\mathbf{x}, \mathbf{x}'')$$

Consequently, if  $\phi$  is a feature map for  $k$ , then

$$\tilde{\phi}(\mathbf{x}') := \left( \mathbf{I} - \frac{\phi(\mathbf{x})\phi(\mathbf{x})^\top}{\sigma^2 + \phi(\mathbf{x})^\top \phi(\mathbf{x})} \right)^{1/2} \phi(\mathbf{x}') = \left( \mathbf{I} - \beta \phi(\mathbf{x})\phi(\mathbf{x})^\top \right) \phi(\mathbf{x}')$$

is a feature map for  $\tilde{k}$ , where

$$\beta := \frac{1}{\sqrt{\sigma^2 + \phi(\mathbf{x})^\top \phi(\mathbf{x})} \left( \sqrt{\sigma^2 + \phi(\mathbf{x})^\top \phi(\mathbf{x})} + \sigma \right)}.$$

**Proof** The kernel update equation follows directly from Eq. (8).

**Step 1: Feature map.** The specified feature map  $\tilde{\phi}$  satisfies

$$\begin{aligned} \tilde{\phi}(\mathbf{x}')^\top \tilde{\phi}(\mathbf{x}'') &= \phi(\mathbf{x}')^\top \left( \mathbf{I} - \frac{\phi(\mathbf{x})\phi(\mathbf{x})^\top}{\sigma^2 + \phi(\mathbf{x})^\top \phi(\mathbf{x})} \right) \phi(\mathbf{x}'') \\ &= k(\mathbf{x}', \mathbf{x}'') - \tilde{k}(\mathbf{x}', \mathbf{x})(\sigma^2 + k(\mathbf{x}, \mathbf{x}))^{-1} k(\mathbf{x}, \mathbf{x}'') \\ &= \tilde{k}(\mathbf{x}', \mathbf{x}''), \end{aligned}$$

hence it is a feature map for  $\tilde{k}$ .

**Step 2: Square root.** According to Lemma D.1, we have

$$\begin{aligned}
 & \left( \mathbf{I} - \frac{\phi(\mathbf{x})\phi(\mathbf{x})^\top}{\sigma^2 + \phi(\mathbf{x})^\top\phi(\mathbf{x})} \right)^{1/2} \\
 &= \mathbf{I} - \frac{1}{(\sigma^2 + \phi(\mathbf{x})^\top\phi(\mathbf{x})) \left( 1 + \sqrt{1 - \phi(\mathbf{x})^\top\phi(\mathbf{x})/(\sigma^2 + \phi(\mathbf{x})^\top\phi(\mathbf{x}))} \right)} \phi(\mathbf{x})\phi(\mathbf{x})^\top \\
 &= \mathbf{I} - \frac{1}{(\sigma^2 + \phi(\mathbf{x})^\top\phi(\mathbf{x}))(1 + \sigma(\sigma^2 + \phi(\mathbf{x})^\top\phi(\mathbf{x}))^{-1/2})} \phi(\mathbf{x})\phi(\mathbf{x})^\top \\
 &= \mathbf{I} - \frac{1}{\sqrt{\sigma^2 + \phi(\mathbf{x})^\top\phi(\mathbf{x})}(\sqrt{\sigma^2 + \phi(\mathbf{x})^\top\phi(\mathbf{x})} + \sigma)} \phi(\mathbf{x})\phi(\mathbf{x})^\top. \quad \blacksquare
 \end{aligned}$$

In our implementation, we keep track of the following quantities:

$$\begin{aligned}
 \Phi_{\mathcal{X}}[\mathbf{x}] &:= \phi_{\rightarrow\text{post}(\mathcal{X}, \sigma^2)}(\mathbf{x}) \\
 c_{\mathcal{X}}[\mathbf{x}] &:= \Phi_{\mathcal{X}}[\mathbf{x}]^\top \Phi_{\mathcal{X}}[\mathbf{x}].
 \end{aligned}$$

Note that  $\phi_{\rightarrow\text{post}(\mathcal{X}, \sigma^2)}$  is not uniquely defined, since any rotation of  $\phi_{\rightarrow\text{post}(\mathcal{X}, \sigma^2)}$  leads to the same kernel. When computing  $\Phi$  as defined above, we do not care which version of  $\phi_{\rightarrow\text{post}(\mathcal{X}, \sigma^2)}$  it corresponds to, as long as the same version of  $\phi_{\rightarrow\text{post}(\mathcal{X}, \sigma^2)}$  is used for all  $\mathbf{x}$ .

Following Proposition D.2,  $\Phi$  and  $c$  can be updated with a new observation  $\mathbf{x} \notin \mathcal{X}$  as follows:

$$\begin{aligned}
 \Phi_{\mathcal{X} \cup \{\mathbf{x}\}}[\mathbf{x}'] &= \Phi_{\mathcal{X}}[\mathbf{x}'] - \beta_{\mathcal{X}}(\mathbf{x}) \Phi_{\mathcal{X}}[\mathbf{x}] \langle \Phi_{\mathcal{X}}[\mathbf{x}], \Phi_{\mathcal{X}}[\mathbf{x}'] \rangle \\
 c_{\mathcal{X} \cup \{\mathbf{x}\}}[\mathbf{x}'] &= \Phi_{\mathcal{X}}[\mathbf{x}']^\top \left( \mathbf{I} - \frac{\Phi_{\mathcal{X}}[\mathbf{x}] \Phi_{\mathcal{X}}[\mathbf{x}]^\top}{\sigma^2 + \Phi_{\mathcal{X}}[\mathbf{x}]^\top \Phi_{\mathcal{X}}[\mathbf{x}]} \right) \Phi_{\mathcal{X}}[\mathbf{x}'] \\
 &= c_{\mathcal{X}}[\mathbf{x}'] - \gamma_{\mathcal{X}}(\mathbf{x})^{-2} \langle \Phi_{\mathcal{X}}[\mathbf{x}], \Phi_{\mathcal{X}}[\mathbf{x}'] \rangle^2,
 \end{aligned}$$

where

$$\begin{aligned}
 \gamma_{\mathcal{X}}(\mathbf{x}) &:= \sqrt{\sigma^2 + c_{\mathcal{X}}[\mathbf{x}]} \\
 \beta_{\mathcal{X}}(\mathbf{x}) &:= \frac{1}{\gamma_{\mathcal{X}}(\mathbf{x})(\gamma_{\mathcal{X}}(\mathbf{x}) + \sigma)}.
 \end{aligned}$$

Together, these considerations lead to a feature-space implementation of MAXDET, presented in Algorithm D.4.

For the complexity analysis of Algorithm D.4, we exclude the computation of the feature matrix  $\phi(\mathcal{X}_{\text{cand}})$ . As for the kernel-space version of MAXDET, the runtime is then dominated by the runtime of ADD, which has a runtime complexity of  $\mathcal{O}(N_{\text{cand}}d_{\text{feat}})$ . Since it is called  $N_{\text{sel}}$  times, the total runtime complexity of Algorithm D.4 is  $\mathcal{O}(N_{\text{cand}}N_{\text{sel}}d_{\text{feat}})$ . If the kernel  $k$  is evaluated by an inner product of the pre-computed features, the runtime of a kernel evaluation scales as  $T_k = \Theta(d_{\text{feat}})$ . In this case, the runtime of the kernel-space version in Algorithm D.3 has a runtime complexity of  $\mathcal{O}(N_{\text{cand}}N_{\text{sel}}(d_{\text{feat}} + N_{\text{sel}}))$ , which is not asymptotically better than the one for Algorithm D.4. However, in our implementation, we observe that the kernel-space version typically runs faster for  $N_{\text{sel}} \lesssim 3d_{\text{feat}}$ , which can be attributed to the smaller constant in the runtime of ADD. It is easily verified that the memory complexity of Algorithm D.4 scales as  $\mathcal{O}(N_{\text{cand}}d_{\text{feat}})$ .



---

**Algorithm D.4** MAXDET pseudocode implementation in feature space for given  $\sigma^2 \geq 0$  using Algorithm D.1.

---

**function** INIT  
 $\Phi \leftarrow \phi(\mathcal{X}_{\text{cand}}) \in \mathbb{R}^{N_{\text{cand}} \times d_{\text{feat}}}$  ▷ feature matrix  
 $\mathbf{c} \leftarrow (\langle \Phi[\mathbf{x}, \cdot]^\top, \Phi[\mathbf{x}, \cdot]^\top \rangle)_{\mathbf{x} \in \mathcal{X}_{\text{cand}}}$  ▷ vector containing the kernel diagonal  
**end function**

**function** ADD( $\mathbf{x}$ )  
 $\gamma \leftarrow \sqrt{\sigma^2 + \mathbf{c}[\mathbf{x}]}$   
 $\beta \leftarrow (\gamma(\gamma + \sigma))^{-1}$   
 $\mathbf{u} \leftarrow \Phi \Phi[\mathbf{x}, \cdot]^\top$   
 $\mathbf{c} \leftarrow \mathbf{c} - \gamma^{-2}(\mathbf{u} \odot \mathbf{u})$   
 $\Phi \leftarrow \Phi - \beta \mathbf{u} \Phi[\mathbf{x}, \cdot]$   
**end function**

**function** NEXT  
**return**  $\operatorname{argmax}_{\mathbf{x} \in \mathcal{X}_{\text{pool}} \setminus \mathcal{X}_{\text{batch}}} \mathbf{c}[\mathbf{x}]$   
**end function**

---

## D.5 BAIT

In this section, we write  $\mathcal{X}_{\text{tp}} := \mathcal{X}_{\text{train}} \cup \mathcal{X}_{\text{pool}}$ .

### D.5.1 CONNECTION BETWEEN KERNEL AND FEATURE MAP FORMULATIONS

We can rewrite BAIT's acquisition function from Eq. (23) as

$$\begin{aligned}
 a(\mathcal{X}) &:= \sum_{\tilde{\mathbf{x}} \in \mathcal{X}_{\text{tp}}} k_{\rightarrow \text{post}(\mathcal{X}, \sigma^2)}(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}) \stackrel{\text{Eq. (10)}}{=} \sigma^2 \sum_{\tilde{\mathbf{x}} \in \mathcal{X}_{\text{tp}}} \phi(\tilde{\mathbf{x}})^\top (\phi(\mathcal{X})^\top \phi(\mathcal{X}) + \sigma^2 \mathbf{I})^{-1} \phi(\tilde{\mathbf{x}}) \\
 &= \sigma^2 \sum_{\tilde{\mathbf{x}} \in \mathcal{X}_{\text{tp}}} \operatorname{tr} \left( \phi(\tilde{\mathbf{x}})^\top (\phi(\mathcal{X})^\top \phi(\mathcal{X}) + \sigma^2 \mathbf{I})^{-1} \phi(\tilde{\mathbf{x}}) \right) \\
 &= \sigma^2 \sum_{\tilde{\mathbf{x}} \in \mathcal{X}_{\text{tp}}} \operatorname{tr} \left( (\phi(\mathcal{X})^\top \phi(\mathcal{X}) + \sigma^2 \mathbf{I})^{-1} \phi(\tilde{\mathbf{x}}) \phi(\tilde{\mathbf{x}})^\top \right) \\
 &= \sigma^2 \operatorname{tr} \left( (\phi(\mathcal{X})^\top \phi(\mathcal{X}) + \sigma^2 \mathbf{I})^{-1} \phi(\mathcal{X}_{\text{tp}})^\top \phi(\mathcal{X}_{\text{tp}}) \right) \\
 &= \sigma^2 \operatorname{tr} \left( (\mathbf{A}_{\mathcal{X}} + \sigma^2 \mathbf{I})^{-1} \mathbf{A}_{\mathcal{X}_{\text{tp}}} \right) ,
 \end{aligned}$$

where

$$\mathbf{A}_{\mathcal{X}} := \phi(\mathcal{X})^\top \phi(\mathcal{X}) .$$

The latter trace-based formulation corresponds to the formulation by Ash et al. (2021).

### D.5.2 FORWARD VERSION

We will first derive an efficient implementation of BAIT-F in feature space, which builds on our derivation of MAXDET in feature space and serves as a basis for BAIT-FB in feature

space. As for the feature space version of MAXDET, we choose to update the features using square roots of rank-1 updates. We will not derive a kernel space version of BAIT here since it appears that a kernel space version would not scale to large data sets. In the following, we will assume  $\sigma^2 > 0$ .

In a single iteration of the BAIT-F selection method, we want to find  $\mathbf{x} \in \mathcal{X}_{\text{pool}} \setminus \mathcal{X}_{\text{batch}}$  *maximizing*

$$\begin{aligned} & a(\mathcal{X}) - a(\mathcal{X} \cup \{\mathbf{x}\}) \\ \stackrel{\text{Proposition D.2}}{=} & \sum_{\tilde{\mathbf{x}} \in \mathcal{X}_{\text{tp}}} \frac{k_{\rightarrow \text{post}}(\mathcal{X}, \sigma^2)(\tilde{\mathbf{x}}, \mathbf{x})^2}{k_{\rightarrow \text{post}}(\mathcal{X}, \sigma^2)(\mathbf{x}, \mathbf{x}) + \sigma^2} \\ = & \frac{\phi_{\rightarrow \text{post}}(\mathcal{X}, \sigma^2)(\mathbf{x})^\top \phi_{\rightarrow \text{post}}(\mathcal{X}, \sigma^2)(\mathcal{X}_{\text{tp}})^\top \phi_{\rightarrow \text{post}}(\mathcal{X}, \sigma^2)(\mathcal{X}_{\text{tp}}) \phi_{\rightarrow \text{post}}(\mathcal{X}, \sigma^2)(\mathbf{x})}{\phi_{\rightarrow \text{post}}(\mathcal{X}, \sigma^2)(\mathbf{x})^\top \phi_{\rightarrow \text{post}}(\mathcal{X}, \sigma^2)(\mathbf{x}) + \sigma^2}. \end{aligned}$$

While the inner product in the denominator of the last expression corresponds to the quantity  $\mathbf{c}_{\mathcal{X}}[\mathbf{x}]$  from the MaxDet feature space implementation, we still need a way to efficiently compute the numerator. Therefore, in addition to the quantities  $\Phi_{\mathcal{X}}$  and  $\mathbf{c}_{\mathcal{X}}$  tracked in the feature-space implementation of MAXDET, we track the following quantities for BAIT:

$$\begin{aligned} \Sigma_{\mathcal{X}} &:= \phi_{\rightarrow \text{post}}(\mathcal{X}, \sigma^2)(\mathcal{X}_{\text{tp}})^\top \phi_{\rightarrow \text{post}}(\mathcal{X}, \sigma^2)(\mathcal{X}_{\text{tp}}) \\ \mathbf{v}_{\mathcal{X}}[\mathbf{x}] &:= \Phi_{\mathcal{X}}[\mathbf{x}]^\top \Sigma_{\mathcal{X}} \Phi_{\mathcal{X}}[\mathbf{x}]. \end{aligned}$$

Using these quantities, we can write

$$a(\mathcal{X}) - a(\mathcal{X} \cup \{\mathbf{x}\}) = \frac{\mathbf{v}_{\mathcal{X}}[\mathbf{x}]}{\mathbf{c}_{\mathcal{X}}[\mathbf{x}] + \sigma^2}.$$

Following Proposition D.2, we obtain the following update equations:

$$\begin{aligned} \Sigma_{\mathcal{X} \cup \{\mathbf{x}\}} &= (\mathbf{I} - \beta_{\mathcal{X}}(\mathbf{x}) \Phi_{\mathcal{X}}[\mathbf{x}] \Phi_{\mathcal{X}}[\mathbf{x}]^\top) \Sigma_{\mathcal{X}} (\mathbf{I} - \beta_{\mathcal{X}}(\mathbf{x}) \Phi_{\mathcal{X}}[\mathbf{x}] \Phi_{\mathcal{X}}[\mathbf{x}]^\top) \\ &= \Sigma_{\mathcal{X}} - \beta_{\mathcal{X}}(\mathbf{x}) \Phi_{\mathcal{X}}[\mathbf{x}] \Phi_{\mathcal{X}}[\mathbf{x}]^\top \Sigma_{\mathcal{X}} - \beta_{\mathcal{X}}(\mathbf{x}) \Sigma_{\mathcal{X}} \Phi_{\mathcal{X}}[\mathbf{x}] \Phi_{\mathcal{X}}[\mathbf{x}]^\top \\ &\quad + \beta_{\mathcal{X}}(\mathbf{x})^2 \Phi_{\mathcal{X}}[\mathbf{x}] \Phi_{\mathcal{X}}[\mathbf{x}]^\top \Sigma_{\mathcal{X}} \Phi_{\mathcal{X}}[\mathbf{x}] \Phi_{\mathcal{X}}[\mathbf{x}]^\top \\ \mathbf{v}_{\mathcal{X} \cup \{\mathbf{x}\}}[\mathbf{x}'] &= \Phi_{\mathcal{X}}[\mathbf{x}']^\top (\mathbf{I} - \gamma_{\mathcal{X}}^{-2}(\mathbf{x}) \Phi_{\mathcal{X}}[\mathbf{x}] \Phi_{\mathcal{X}}[\mathbf{x}]^\top) \Sigma_{\mathcal{X}} (\mathbf{I} - \gamma_{\mathcal{X}}(\mathbf{x})^{-2} \Phi_{\mathcal{X}}[\mathbf{x}] \Phi_{\mathcal{X}}[\mathbf{x}]^\top) \Phi_{\mathcal{X}}[\mathbf{x}'] \\ &= \mathbf{v}_{\mathcal{X}}[\mathbf{x}'] - 2\gamma_{\mathcal{X}}(\mathbf{x})^{-2} \Phi_{\mathcal{X}}[\mathbf{x}']^\top \Phi_{\mathcal{X}}[\mathbf{x}] \Phi_{\mathcal{X}}[\mathbf{x}]^\top \Sigma_{\mathcal{X}} \Phi_{\mathcal{X}}[\mathbf{x}'] \\ &\quad + \gamma_{\mathcal{X}}(\mathbf{x})^{-4} \Phi_{\mathcal{X}}[\mathbf{x}']^\top \Phi_{\mathcal{X}}[\mathbf{x}] \Phi_{\mathcal{X}}[\mathbf{x}]^\top \Sigma_{\mathcal{X}} \Phi_{\mathcal{X}}[\mathbf{x}] \Phi_{\mathcal{X}}[\mathbf{x}]^\top \Phi_{\mathcal{X}}[\mathbf{x}']. \end{aligned}$$

This leads to the pseudocode in Algorithm D.5. For the runtime analysis, we neglect the time for evaluating  $\phi$  as usual. Then, the runtime of INIT is  $\mathcal{O}((N_{\text{train}} + N_{\text{pool}})d_{\text{feat}}^2)$ , the runtime of ADD is  $\mathcal{O}((N_{\text{cand}} + d_{\text{feat}})d_{\text{feat}})$  and the runtime of NEXT is  $\mathcal{O}(N_{\text{pool}})$ . Hence, the overall runtime of BAIT-F is  $\mathcal{O}(N_{\text{cand}}N_{\text{sel}}d_{\text{feat}} + (N_{\text{train}} + N_{\text{pool}})d_{\text{feat}}^2)$ . The memory complexity is  $\mathcal{O}((N_{\text{cand}} + d_{\text{feat}})d_{\text{feat}})$ .

### D.5.3 FORWARD-BACKWARD VERSION

To fit BAIT-FB into our framework, we first extend our iterative selection template to include a backward selection step. The extended template is shown in Algorithm D.6. Here, we have an additional parameter  $N_{\text{extra}}$  specifying how many additional batch elements will be

---

**Algorithm D.5** BAIT-F pseudocode implementation in feature space for given  $\sigma^2 > 0$  using Algorithm D.1.

---

**function** INIT

$\Phi \leftarrow \phi(\mathcal{X}_{\text{cand}}) \in \mathbb{R}^{N_{\text{cand}} \times d_{\text{feat}}}$  ▷ feature matrix

$\mathbf{c} \leftarrow (\langle \Phi[\mathbf{x}, \cdot]^\top, \Phi[\mathbf{x}, \cdot]^\top \rangle)_{\mathbf{x} \in \mathcal{X}_{\text{cand}}}$  ▷ vector containing the kernel diagonal

$\Sigma \leftarrow \phi(\mathcal{X}_{\text{tp}})^\top \phi(\mathcal{X}_{\text{tp}})$  ▷ Train and pool second moment matrix

$\mathbf{v} \leftarrow (\langle \Phi[\mathbf{x}, \cdot]^\top, \Sigma \Phi[\mathbf{x}, \cdot]^\top \rangle)_{\mathbf{x} \in \mathcal{X}_{\text{cand}}}$  ▷ Numerator of the acquisition function

**end function**

**function** ADD( $\mathbf{x}$ )

$\gamma \leftarrow \sqrt{\sigma^2 + \mathbf{c}[\mathbf{x}]}$

$\beta \leftarrow (\gamma(\gamma + \sigma))^{-1}$

$\mathbf{u} \leftarrow \Phi \Phi[\mathbf{x}, \cdot]^\top$

$\tilde{\mathbf{u}} \leftarrow \mathbf{u} \odot \mathbf{u}$

$\mathbf{w} \leftarrow \Sigma \Phi[\mathbf{x}, \cdot]^\top$

$\tilde{\mathbf{v}} \leftarrow \mathbf{v}[\mathbf{x}]$

$\mathbf{v} \leftarrow \mathbf{v} - 2\gamma^{-2}(\Phi \mathbf{w}) \odot \mathbf{u} + \gamma^{-4} \tilde{\mathbf{v}} \tilde{\mathbf{u}}$

$\mathbf{A} \leftarrow \mathbf{w} \Phi[\mathbf{x}, \cdot]$

$\Sigma \leftarrow \Sigma - \beta(\mathbf{A} + \mathbf{A}^\top) + \beta^2 \tilde{\mathbf{v}} \Phi[\mathbf{x}, \cdot]^\top \Phi[\mathbf{x}, \cdot]$

$\mathbf{c} \leftarrow \mathbf{c} - \gamma^{-2} \tilde{\mathbf{u}}$

$\Phi \leftarrow \Phi - \beta \mathbf{u} \Phi[\mathbf{x}, \cdot]$

**end function**

**function** NEXT

**return**  $\operatorname{argmax}_{\mathbf{x} \in \mathcal{X}_{\text{pool}} \setminus \mathcal{X}_{\text{batch}}} \frac{\mathbf{v}[\mathbf{x}]}{\sigma^2 + \mathbf{c}[\mathbf{x}]}$

**end function**

---

selected in the forward step and then removed in the backward step. Following Ash et al. (2021), we set  $N_{\text{extra}} := \min\{N_{\text{batch}}, N_{\text{pool}} - N_{\text{batch}}\}$  in our experiments.

To implement BAIT-FB within Algorithm D.6, we can reuse the methods INIT, ADD and NEXT from BAIT-F. In addition, we need to implement the methods REMOVE and NEXTBACKWARD for the backward step. To this end, the following proposition shows how the kernel and feature map update when *removing* a point  $\mathbf{x}$  from the set  $\mathcal{X}$  of observed points:

**Proposition D.3** (Backward update). *For a kernel  $k$  and  $\sigma^2 > 0$ , let  $\tilde{k} := k_{\rightarrow \text{post}(\{\mathbf{x}\}, \sigma^2)}$ . Then,*

$$k(\mathbf{x}', \mathbf{x}'') = \tilde{k}(\mathbf{x}', \mathbf{x}'') + \tilde{k}(\mathbf{x}', \mathbf{x})(\sigma^2 - \tilde{k}(\mathbf{x}, \mathbf{x}))^{-1} \tilde{k}(\mathbf{x}, \mathbf{x}'') .$$

Consequently, if  $\tilde{\phi}$  is a feature map for  $\tilde{k}$ , then

$$\phi(\mathbf{x}') := \left( \mathbf{I} + \frac{\tilde{\phi}(\mathbf{x}) \tilde{\phi}(\mathbf{x})^\top}{\sigma^2 - \tilde{\phi}(\mathbf{x})^\top \tilde{\phi}(\mathbf{x})} \right)^{1/2} \tilde{\phi}(\mathbf{x}') = \left( \mathbf{I} + \tilde{\beta} \tilde{\phi}(\mathbf{x}) \tilde{\phi}(\mathbf{x})^\top \right) \tilde{\phi}(\mathbf{x}')$$

---

**Algorithm D.6** Forward-backward selection algorithm template involving five customizable functions INIT, ADD, NEXT, REMOVE, NEXTBACKWARD that are allowed to have side effects (i.e., read/write variables in SELECT).

---

```

function SELECT( $k$ ,  $\mathcal{X}_{\text{train}}$ ,  $\mathcal{X}_{\text{pool}}$ ,  $N_{\text{batch}}$ ,  $N_{\text{extra}}$ , mode  $\in$  {P, TP})
     $\mathcal{X}_{\text{mode}} \leftarrow \mathcal{X}_{\text{train}}$  if mode = TP else  $\emptyset$ 
     $\mathcal{X}_{\text{cand}} := \mathcal{X}_{\text{mode}} \cup \mathcal{X}_{\text{pool}}$ 
     $\mathcal{X}_{\text{batch}} \leftarrow \emptyset$ 
    INIT
    for  $\mathbf{x}$  in  $\mathcal{X}_{\text{mode}}$  do
        ADD( $\mathbf{x}$ )
    end for
    for  $i$  from 1 to  $N_{\text{batch}} + N_{\text{extra}}$  do
         $\mathbf{x} \leftarrow$  NEXT
        if  $\mathbf{x} \in \mathcal{X}_{\text{batch}} \cup \mathcal{X}_{\text{train}}$  (failed selection) then
            ensure  $|\mathcal{X}_{\text{batch}}| = N_{\text{batch}}$  by removing the latest samples or filling up with random samples
            return  $\mathcal{X}_{\text{batch}}$ 
        end if
         $\mathcal{X}_{\text{batch}} \leftarrow \mathcal{X}_{\text{batch}} \cup \{\mathbf{x}\}$ 
        ADD( $\mathbf{x}$ )
    end for
    for  $i$  from 1 to  $N_{\text{extra}}$  do
         $\mathbf{x} \leftarrow$  NEXTBACKWARD
        if  $\mathbf{x} \notin \mathcal{X}_{\text{batch}} \cup \mathcal{X}_{\text{train}}$  (failed selection) then
            ensure  $|\mathcal{X}_{\text{batch}}| = N_{\text{batch}}$  by removing the latest samples
            return  $\mathcal{X}_{\text{batch}}$ 
        end if
         $\mathcal{X}_{\text{batch}} \leftarrow \mathcal{X}_{\text{batch}} \setminus \{\mathbf{x}\}$ 
        REMOVE( $\mathbf{x}$ )
    end for
    return  $\mathcal{X}_{\text{batch}}$ 
end function
    
```

---

is a feature map for  $k$ , where

$$\tilde{\beta} := \frac{1}{\sqrt{\sigma^2 - \tilde{\phi}(\mathbf{x})^\top \tilde{\phi}(\mathbf{x})} \left( \sqrt{\sigma^2 - \tilde{\phi}(\mathbf{x})^\top \tilde{\phi}(\mathbf{x})} + \sigma \right)}.$$

**Proof Step 1: Finding  $k(\mathbf{x}, \mathbf{x})$ .** We have

$$\tilde{k}(\mathbf{x}, \mathbf{x}) = k(\mathbf{x}, \mathbf{x}) - \frac{k(\mathbf{x}, \mathbf{x})^2}{k(\mathbf{x}, \mathbf{x}) + \sigma^2} = \frac{k(\mathbf{x}, \mathbf{x})\sigma^2}{k(\mathbf{x}, \mathbf{x}) + \sigma^2} < \sigma^2.$$

Hence,

$$\sigma^2 - \tilde{k}(\mathbf{x}, \mathbf{x}) = \frac{\sigma^2(k(\mathbf{x}, \mathbf{x}) + \sigma^2)}{k(\mathbf{x}, \mathbf{x}) + \sigma^2} - \frac{k(\mathbf{x}, \mathbf{x})\sigma^2}{k(\mathbf{x}, \mathbf{x}) + \sigma^2} = \frac{\sigma^4}{k(\mathbf{x}, \mathbf{x}) + \sigma^2},$$

which yields

$$\frac{\sigma^2}{\sigma^2 - \tilde{k}(\mathbf{x}, \mathbf{x})} = \frac{k(\mathbf{x}, \mathbf{x}) + \sigma^2}{\sigma^2}.$$

**Step 2: Finding  $k(\mathbf{x}', \mathbf{x})$ .** Now, we compute

$$\tilde{k}(\mathbf{x}', \mathbf{x}) = k(\mathbf{x}', \mathbf{x}) - k(\mathbf{x}', \mathbf{x}) \frac{k(\mathbf{x}, \mathbf{x})}{k(\mathbf{x}, \mathbf{x}) + \sigma^2} = k(\mathbf{x}', \mathbf{x}) \frac{\sigma^2}{k(\mathbf{x}, \mathbf{x}) + \sigma^2},$$

which yields

$$k(\mathbf{x}', \mathbf{x}) = \tilde{k}(\mathbf{x}', \mathbf{x}) \frac{k(\mathbf{x}, \mathbf{x}) + \sigma^2}{\sigma^2} = \tilde{k}(\mathbf{x}', \mathbf{x}) \frac{\sigma^2}{\sigma^2 - \tilde{k}(\mathbf{x}, \mathbf{x})}.$$

**Step 3: Finding  $k(\mathbf{x}', \mathbf{x}'')$ .** Finally, we have

$$\tilde{k}(\mathbf{x}', \mathbf{x}'') = k(\mathbf{x}', \mathbf{x}'') - \frac{k(\mathbf{x}', \mathbf{x})k(\mathbf{x}, \mathbf{x}'')}{k(\mathbf{x}, \mathbf{x}) + \sigma^2},$$

which yields

$$\begin{aligned} k(\mathbf{x}', \mathbf{x}'') &= \tilde{k}(\mathbf{x}', \mathbf{x}'') + k(\mathbf{x}', \mathbf{x}) \cdot \frac{1}{k(\mathbf{x}, \mathbf{x}) + \sigma^2} \cdot k(\mathbf{x}, \mathbf{x}'') \\ &= \tilde{k}(\mathbf{x}', \mathbf{x}'') + \tilde{k}(\mathbf{x}', \mathbf{x}) \frac{k(\mathbf{x}, \mathbf{x}) + \sigma^2}{\sigma^2} \cdot \frac{1}{k(\mathbf{x}, \mathbf{x}) + \sigma^2} \cdot \tilde{k}(\mathbf{x}, \mathbf{x}'') \frac{\sigma^2}{\sigma^2 - \tilde{k}(\mathbf{x}, \mathbf{x})} \\ &= \tilde{k}(\mathbf{x}', \mathbf{x}'') + \frac{\tilde{k}(\mathbf{x}', \mathbf{x})\tilde{k}(\mathbf{x}, \mathbf{x}'')}{\sigma^2 - \tilde{k}(\mathbf{x}, \mathbf{x})}. \end{aligned}$$

**Step 4: Feature map.** The specified feature map  $\phi$  satisfies

$$\begin{aligned} \phi(\mathbf{x}')^\top \phi(\mathbf{x}'') &= \tilde{\phi}(\mathbf{x}')^\top \left( \mathbf{I} + \frac{\tilde{\phi}(\mathbf{x})\tilde{\phi}(\mathbf{x})^\top}{\sigma^2 - \tilde{\phi}(\mathbf{x})^\top \tilde{\phi}(\mathbf{x})} \right) \tilde{\phi}(\mathbf{x}'') \\ &= \tilde{k}(\mathbf{x}', \mathbf{x}'') + \tilde{k}(\mathbf{x}', \mathbf{x})(\sigma^2 - \tilde{k}(\mathbf{x}, \mathbf{x}))^{-1}\tilde{k}(\mathbf{x}, \mathbf{x}'') \end{aligned}$$

$$= k(\mathbf{x}', \mathbf{x}'') ,$$

hence it is a feature map for  $k$ .

**Step 5: Square root.** According to Lemma D.1, we have

$$\begin{aligned} & \left( \mathbf{I} + \frac{\tilde{\phi}(\mathbf{x})\tilde{\phi}(\mathbf{x})^\top}{\sigma^2 - \tilde{\phi}(\mathbf{x})^\top\tilde{\phi}(\mathbf{x})} \right)^{1/2} \\ &= \mathbf{I} + \frac{1}{(\sigma^2 - \tilde{\phi}(\mathbf{x})^\top\tilde{\phi}(\mathbf{x})) \left( 1 + \sqrt{1 + \tilde{\phi}(\mathbf{x})^\top\tilde{\phi}(\mathbf{x})/(\sigma^2 - \tilde{\phi}(\mathbf{x})^\top\tilde{\phi}(\mathbf{x}))} \right)} \tilde{\phi}(\mathbf{x})\tilde{\phi}(\mathbf{x})^\top \\ &= \mathbf{I} + \frac{1}{(\sigma^2 - \tilde{\phi}(\mathbf{x})^\top\tilde{\phi}(\mathbf{x}))(1 + \sigma(\sigma^2 - \tilde{\phi}(\mathbf{x})^\top\tilde{\phi}(\mathbf{x}))^{-1/2})} \tilde{\phi}(\mathbf{x})\tilde{\phi}(\mathbf{x})^\top \\ &= \mathbf{I} + \frac{1}{\sqrt{\sigma^2 - \tilde{\phi}(\mathbf{x})^\top\tilde{\phi}(\mathbf{x})}(\sqrt{\sigma^2 - \tilde{\phi}(\mathbf{x})^\top\tilde{\phi}(\mathbf{x})} + \sigma)} \tilde{\phi}(\mathbf{x})\tilde{\phi}(\mathbf{x})^\top . \quad \blacksquare \end{aligned}$$

To formulate the backward update for our tracked quantities  $\Phi, \mathbf{c}, \Sigma$  and  $\mathbf{v}$ , we define for  $\mathbf{x} \in \mathcal{X}$  the scalars

$$\begin{aligned} \tilde{\gamma}_{\mathcal{X}}(\mathbf{x}) &:= \sqrt{\sigma^2 - \mathbf{c}[\mathbf{x}]} \\ \tilde{\beta}_{\mathcal{X}}(\mathbf{x}) &:= \frac{1}{\gamma(\gamma + \sigma)} . \end{aligned}$$

Then, using Proposition D.3, we can compute the backwards update as

$$\begin{aligned} \Phi_{\mathcal{X} \setminus \{\mathbf{x}\}}[\mathbf{x}'] &= \Phi_{\mathcal{X}}[\mathbf{x}'] + \tilde{\beta}_{\mathcal{X}}(\mathbf{x})\Phi_{\mathcal{X}}[\mathbf{x}]\langle \Phi_{\mathcal{X}}[\mathbf{x}], \Phi_{\mathcal{X}}[\mathbf{x}'] \rangle \\ \mathbf{c}_{\mathcal{X} \setminus \{\mathbf{x}\}}[\mathbf{x}'] &= \Phi_{\mathcal{X}}[\mathbf{x}']^\top \left( \mathbf{I} + \frac{\Phi_{\mathcal{X}}[\mathbf{x}]\Phi_{\mathcal{X}}[\mathbf{x}]^\top}{\sigma^2 - \Phi_{\mathcal{X}}[\mathbf{x}]^\top\Phi_{\mathcal{X}}[\mathbf{x}]} \right) \Phi_{\mathcal{X}}[\mathbf{x}'] \\ &= \mathbf{c}_{\mathcal{X}}[\mathbf{x}'] + \tilde{\gamma}_{\mathcal{X}}(\mathbf{x})^{-2} \langle \Phi_{\mathcal{X}}[\mathbf{x}], \Phi_{\mathcal{X}}[\mathbf{x}'] \rangle^2 \\ \Sigma_{\mathcal{X} \setminus \{\mathbf{x}\}} &= (\mathbf{I} - \beta_{\mathcal{X}}(\mathbf{x})\Phi_{\mathcal{X}}[\mathbf{x}]\Phi_{\mathcal{X}}[\mathbf{x}]^\top)\Sigma_{\mathcal{X}}(\mathbf{I} - \beta_{\mathcal{X}}(\mathbf{x})\Phi_{\mathcal{X}}[\mathbf{x}]\Phi_{\mathcal{X}}[\mathbf{x}]^\top) \\ &= \Sigma_{\mathcal{X}} + \tilde{\beta}_{\mathcal{X}}(\mathbf{x})\Phi_{\mathcal{X}}[\mathbf{x}]\Phi_{\mathcal{X}}[\mathbf{x}]^\top\Sigma_{\mathcal{X}} + \tilde{\beta}_{\mathcal{X}}(\mathbf{x})\Sigma_{\mathcal{X}}\Phi_{\mathcal{X}}[\mathbf{x}]\Phi_{\mathcal{X}}[\mathbf{x}]^\top \\ &\quad + \beta_{\mathcal{X}}(\mathbf{x})^2\Phi_{\mathcal{X}}[\mathbf{x}]\Phi_{\mathcal{X}}[\mathbf{x}]^\top\Sigma_{\mathcal{X}}\Phi_{\mathcal{X}}[\mathbf{x}]\Phi_{\mathcal{X}}[\mathbf{x}]^\top \\ \mathbf{v}_{\mathcal{X} \setminus \{\mathbf{x}\}}[\mathbf{x}'] &= \Phi_{\mathcal{X}}[\mathbf{x}']^\top (\mathbf{I} + \tilde{\gamma}_{\mathcal{X}}^{-2}(\mathbf{x})\Phi_{\mathcal{X}}[\mathbf{x}]\Phi_{\mathcal{X}}[\mathbf{x}]^\top)\Sigma_{\mathcal{X}}(\mathbf{I} + \tilde{\gamma}_{\mathcal{X}}(\mathbf{x})^{-2}\Phi_{\mathcal{X}}[\mathbf{x}]\Phi_{\mathcal{X}}[\mathbf{x}]^\top)\Phi_{\mathcal{X}}[\mathbf{x}'] \\ &= \mathbf{v}_{\mathcal{X}}[\mathbf{x}'] + 2\tilde{\gamma}_{\mathcal{X}}(\mathbf{x})^{-2}\Phi_{\mathcal{X}}[\mathbf{x}']^\top\Phi_{\mathcal{X}}[\mathbf{x}]\Phi_{\mathcal{X}}[\mathbf{x}]^\top\Sigma_{\mathcal{X}}\Phi_{\mathcal{X}}[\mathbf{x}'] \\ &\quad + \tilde{\gamma}_{\mathcal{X}}(\mathbf{x})^{-4}\Phi_{\mathcal{X}}[\mathbf{x}']^\top\Phi_{\mathcal{X}}[\mathbf{x}]\Phi_{\mathcal{X}}[\mathbf{x}]^\top\Sigma_{\mathcal{X}}\Phi_{\mathcal{X}}[\mathbf{x}]\Phi_{\mathcal{X}}[\mathbf{x}]^\top\Phi_{\mathcal{X}}[\mathbf{x}'] . \end{aligned}$$

For the backward step, we want to find  $\mathbf{x} \in \mathcal{X}$  *minimizing*

$$\begin{aligned} & a(\mathcal{X} \setminus \{\mathbf{x}\}) - a(\mathcal{X}) \\ \stackrel{\text{Proposition D.3}}{=} & \sum_{\tilde{\mathbf{x}} \in \mathcal{X}_{\text{tp}}} \frac{k_{\rightarrow \text{post}(\mathcal{X}, \sigma^2)}(\tilde{\mathbf{x}}, \mathbf{x})^2}{\sigma^2 - k_{\rightarrow \text{post}(\mathcal{X}, \sigma^2)}(\mathbf{x}, \mathbf{x})} \\ = & \frac{\phi_{\rightarrow \text{post}(\mathcal{X}, \sigma^2)}(\mathbf{x})^\top \phi_{\rightarrow \text{post}(\mathcal{X}, \sigma^2)}(\mathcal{X}_{\text{tp}})^\top \phi_{\rightarrow \text{post}(\mathcal{X}, \sigma^2)}(\mathcal{X}_{\text{tp}}) \phi_{\rightarrow \text{post}(\mathcal{X}, \sigma^2)}(\mathbf{x})}{\sigma^2 - \phi_{\rightarrow \text{post}(\mathcal{X}, \sigma^2)}(\mathbf{x})^\top \phi_{\rightarrow \text{post}(\mathcal{X}, \sigma^2)}(\mathbf{x})} \\ = & \frac{\mathbf{v}_{\mathcal{X}}[\mathbf{x}]}{\sigma^2 - \mathbf{c}_{\mathcal{X}}[\mathbf{x}]} . \end{aligned}$$

---

**Algorithm D.7** Functions REMOVE and NEXTBACKWARD that, together with Algorithm D.5 and Algorithm D.6, yield a pseudocode implementation of BAIT-FB in feature space for given  $\sigma^2 > 0$ .

---

```

function REMOVE( $\mathbf{x}$ )
     $\tilde{\gamma} \leftarrow \sqrt{\sigma^2 - \mathbf{c}[\mathbf{x}]}$ 
     $\tilde{\beta} \leftarrow (\gamma(\gamma + \sigma))^{-1}$ 
     $\mathbf{u} \leftarrow \Phi\Phi[\mathbf{x}, \cdot]^\top$ 
     $\tilde{\mathbf{u}} \leftarrow \mathbf{u} \odot \mathbf{u}$ 
     $\mathbf{w} \leftarrow \Sigma\Phi[\mathbf{x}, \cdot]^\top$ 
     $\tilde{v} \leftarrow \mathbf{v}[\mathbf{x}]$ 
     $\mathbf{v} \leftarrow \mathbf{v} + 2\tilde{\gamma}^{-2}(\Phi\mathbf{w}) \odot \mathbf{u} + \tilde{\gamma}^{-4}\tilde{v}\tilde{\mathbf{u}}$ 
     $\mathbf{A} \leftarrow \mathbf{w}\Phi[\mathbf{x}, \cdot]$ 
     $\Sigma \leftarrow \Sigma + \tilde{\beta}(\mathbf{A} + \mathbf{A}^\top) + \tilde{\beta}^2\tilde{v}\Phi[\mathbf{x}, \cdot]^\top\Phi[\mathbf{x}, \cdot]$ 
     $\mathbf{c} \leftarrow \mathbf{c} + \tilde{\gamma}^{-2}\tilde{\mathbf{u}}$ 
     $\Phi \leftarrow \Phi + \tilde{\beta}\mathbf{u}\Phi[\mathbf{x}, \cdot]$ 
end function

function NEXTBACKWARD
    return  $\operatorname{argmin}_{\mathbf{x} \in \mathcal{X}_{\text{batch}}} \frac{\mathbf{v}[\mathbf{x}]}{\sigma^2 - \mathbf{c}[\mathbf{x}]}$ 
end function
    
```

---

The corresponding implementation of REMOVE and NEXTBACKWARD is given in Algorithm D.7, completing the implementation of BAIT-FB. The runtimes of REMOVE and NEXTBACKWARD are equivalent to those of ADD and NEXT, respectively, since the implementation is almost identical. Hence, the runtime complexity of BAIT-FB is given by  $\mathcal{O}(N_{\text{cand}}(N_{\text{sel}} + 2N_{\text{extra}})d_{\text{feat}} + (N_{\text{train}} + N_{\text{pool}})d_{\text{feat}}^2)$ . The memory complexity is again  $\mathcal{O}((N_{\text{cand}} + d_{\text{feat}})d_{\text{feat}})$ .

## D.6 FRANKWOLFE

Pinsler et al. (2019) proposed to apply the Frank-Wolfe constrained optimization algorithm (Frank and Wolfe, 1956) to the problem of sparsely approximating the empirical kernel mean embedding of  $\mathcal{X}_{\text{pool}}$  with non-negative weights. Like MAXDET, the resulting FRANKWOLFE method allows for a kernel-space and a feature-space implementation. Pinsler et al. (2019) used both versions in their experiments and presented the kernel-space version as pseudocode. Algorithm D.8 is an optimized adaptation of their kernel-space version to our framework. A difference between our version and theirs is that in our version, NEXT does not allow choosing a previously selected point. Hence, our version prevents the possibility of generating smaller batches by selecting the same point multiple times. Moreover, our version reduces the runtime complexity from  $\mathcal{O}(N_{\text{cand}}^2(T_k + N_{\text{sel}}))$  to  $\mathcal{O}(N_{\text{cand}}^2(T_k + 1))$  by reusing previously computed quantities in ADD. The memory complexity of Algorithm D.8 is  $\mathcal{O}(N_{\text{cand}}^2)$ , which can be reduced to  $\mathcal{O}(N_{\text{cand}})$  by not storing the kernel matrix  $\mathbf{K}$ , at the cost of having to recompute some kernel values in ADD.

---

**Algorithm D.8** FRANKWOLFE pseudocode implementation in kernel space using Algorithm D.1, following Pinsler et al. (2019).

---

**function** INIT

$\mathbf{K} \leftarrow (k(\mathbf{x}, \tilde{\mathbf{x}}))_{\mathbf{x}, \tilde{\mathbf{x}} \in \mathcal{X}_{\text{cand}}}$  ▷ Corresponds to  $\langle \mathcal{L}_n, \mathcal{L}_n \rangle$   
 $\mathbf{c} \leftarrow (\sqrt{\mathbf{K}}[\mathbf{x}, \mathbf{x}])_{\mathbf{x} \in \mathcal{X}_{\text{cand}}}$  ▷ Corresponds to  $\sigma_n$   
 $r \leftarrow \sum_{\mathbf{x} \in \mathcal{X}_{\text{cand}}} \mathbf{c}[\mathbf{x}]$  ▷ Corresponds to  $\sigma$   
 $\mathbf{u} \leftarrow (\sum_{\tilde{\mathbf{x}} \in \mathcal{X}_{\text{cand}}} \mathbf{K}[\mathbf{x}, \tilde{\mathbf{x}}])_{\mathbf{x} \in \mathcal{X}_{\text{cand}}}$  ▷ Corresponds to  $\langle \mathcal{L}, \mathcal{L}_n \rangle$   
 $\mathbf{v} \leftarrow (0)_{\mathbf{x} \in \mathcal{X}_{\text{cand}}}$  ▷ Corresponds to  $\langle \mathcal{L}(\mathbf{w}), \mathcal{L}_n \rangle$   
 $s \leftarrow 0$  ▷ Corresponds to  $\langle \mathcal{L}(\mathbf{w}), \mathcal{L}(\mathbf{w}) \rangle$   
 $t \leftarrow 0$  ▷ Corresponds to  $\langle \mathcal{L}(\mathbf{w}), \mathcal{L} \rangle$

**end function**

**function** ADD( $\mathbf{x}$ )

$\gamma \leftarrow \frac{r\mathbf{c}[\mathbf{x}]^{-1}(\mathbf{u}[\mathbf{x}] - \mathbf{v}[\mathbf{x}]) + s - t}{r^2 - 2r\mathbf{c}[\mathbf{x}]^{-1}\mathbf{v}[\mathbf{x}] + s}$   
 $s \leftarrow (1 - \gamma)^2 s + 2(1 - \gamma)\gamma r\mathbf{c}[\mathbf{x}]^{-1}\mathbf{v}[\mathbf{x}] + \gamma^2 r^2$   
 $t \leftarrow (1 - \gamma)t + \gamma r\mathbf{c}[\mathbf{x}]^{-1}\mathbf{u}[\mathbf{x}]$   
 $\mathbf{v} \leftarrow (1 - \gamma)\mathbf{v} + \gamma r\mathbf{c}[\mathbf{x}]^{-1}\mathbf{K}[\mathbf{x}, \cdot]$

**end function**

**function** NEXT

**return**  $\operatorname{argmax}_{\mathbf{x} \in \mathcal{X}_{\text{pool}} \setminus \mathcal{X}_{\text{batch}}} \mathbf{c}[\mathbf{x}]^{-1}(\mathbf{u}[\mathbf{x}] - \mathbf{v}[\mathbf{x}])$

**end function**

---

The quadratic complexity in  $N_{\text{cand}}$  for the kernel-space version of FRANKWOLFE shown in Algorithm D.8 makes it infeasible for large  $N_{\text{pool}}$ , such as in our experiments. However, FRANKWOLFE can be realized much more efficiently in moderate-dimensional feature spaces, as shown in Algorithm D.9 and implemented in our code and (less efficiently) in the code of Pinsler et al. (2019). In Algorithm D.9, when ignoring the computation of  $\Phi$ , the runtime complexity of INIT is  $\mathcal{O}(N_{\text{cand}}d_{\text{feat}})$ , the runtime complexity of ADD is  $\mathcal{O}(d_{\text{feat}})$ , and the runtime of NEXT is  $\mathcal{O}(N_{\text{pool}}d_{\text{feat}})$ . In total, we obtain a runtime complexity of  $\mathcal{O}((N_{\text{cand}} + N_{\text{pool}}N_{\text{batch}} + N_{\text{sel}})d_{\text{feat}}) = \mathcal{O}((N_{\text{cand}} + N_{\text{pool}}N_{\text{batch}})d_{\text{feat}})$ . The memory complexity of Algorithm D.9 is  $\mathcal{O}(N_{\text{cand}}d_{\text{feat}})$ .

## D.7 MAXDIST

The MAXDIST selection method has been proposed various times in the literature under many different names. Up to the selection of the first two points, it is equivalent to the Kennard-Stone algorithm (Kennard and Stone, 1969) proposed for experimental design. Rosenkrantz et al. (1977) proposed it under the name *farthest insertion* to generate an insertion order for constructing an approximate TSP solution. Later, Gonzalez (1985) proposed it as an approximation algorithm for a clustering problem. In this context, it is also known as *farthest-point clustering* (Bern and Eppstein, 1996) or *k-center greedy* (Sener and Savarese, 2018). Moreover, it has been proposed as an initialization method for k-means clustering (Katsavounidis et al., 1994). MAXDIST is also equivalent to the *geometric greedy*



---

**Algorithm D.9** FRANKWOLFE pseudocode implementation in feature space using Algorithm D.1.

---

**function** INIT

$\Phi \leftarrow \phi(\mathcal{X}_{\text{cand}}) \in \mathbb{R}^{N_{\text{cand}} \times d_{\text{feat}}}$	▷ Corresponds to $\mathcal{L}_n$
$\mathbf{c} \leftarrow (\ \Phi[\mathbf{x}, \cdot]\ _2)_{\mathbf{x} \in \mathcal{X}_{\text{cand}}}$	▷ Corresponds to $\sigma_n$
$r \leftarrow \sum_{\mathbf{x} \in \mathcal{X}_{\text{cand}}} \mathbf{c}[\mathbf{x}]$	▷ Corresponds to $\sigma$
$\tilde{\Phi} \leftarrow (\mathbf{c}[\mathbf{x}]^{-1} \Phi[\mathbf{x}, i])_{\mathbf{x} \in \mathcal{X}_{\text{cand}}, i \in \{1, \dots, d_{\text{feat}}\}} \in \mathbb{R}^{N_{\text{cand}} \times d_{\text{feat}}}$	▷ Corresponds to $\frac{1}{\sigma_n} \mathcal{L}_n$
$\mathbf{u} \leftarrow \sum_{\mathbf{x} \in \mathcal{X}_{\text{cand}}} \tilde{\Phi}[\mathbf{x}, \cdot]$	▷ Corresponds to $\mathcal{L}$
$\mathbf{v} \leftarrow \mathbf{0} \in \mathbb{R}^{d_{\text{feat}}}$	▷ Corresponds to $\mathcal{L}(\mathbf{w})$

**end function**

**function** ADD( $\mathbf{x}$ )

$\gamma \leftarrow \frac{\langle r \tilde{\Phi}[\mathbf{x}, \cdot] - \mathbf{v}, \mathbf{u} - \mathbf{v} \rangle}{\langle r \tilde{\Phi}[\mathbf{x}, \cdot] - \mathbf{v}, r \tilde{\Phi}[\mathbf{x}, \cdot] - \mathbf{v} \rangle}$
$\mathbf{v} \leftarrow (1 - \gamma)\mathbf{v} + \gamma r \tilde{\Phi}[\mathbf{x}, \cdot]$

**end function**

**function** NEXT

<b>return</b> $\operatorname{argmax}_{\mathbf{x} \in \mathcal{X}_{\text{pool}} \setminus \mathcal{X}_{\text{batch}}} \langle \tilde{\Phi}[\mathbf{x}, \cdot], \mathbf{u} - \mathbf{v} \rangle$
--

**end function**

---

algorithm for kernel interpolation (De Marchi et al., 2005). When used with  $N_{\text{batch}} = N_{\text{pool}}$  to construct an ordering of the points, it is known as *farthest-first traversal* or *greedy permutation* of a finite metric space (Eppstein et al., 2020).

Algorithm D.10 shows a pseudocode implementation of MAXDIST. Since ADD has a runtime of  $\mathcal{O}(N_{\text{pool}}(T_k + 1))$ , the runtime of Algorithm D.10 is  $\mathcal{O}(N_{\text{pool}}N_{\text{sel}}(T_k + 1))$ . The memory complexity is  $\mathcal{O}(N_{\text{pool}})$ .

We will now investigate approximation guarantees for MAXDIST with respect to a covering objective, called *minmax radius clustering* or *euclidean  $k$ -center problem* (Bern and Eppstein, 1996). Approximation guarantees can also be given for a related objective called *minmax diameter clustering* (Gonzalez, 1985; Bern and Eppstein, 1996), which will not be discussed here. The following notation will help to define the minmax radius clustering problem:

**Definition D.4.** For a given pseudometric  $d$  (i.e., a metric except that  $d(\mathbf{x}, \mathbf{x}') = 0$  is allowed for  $\mathbf{x} \neq \mathbf{x}'$ ), batch size  $N_{\text{batch}} \in \mathbb{N}$  and batch  $\mathcal{X}_{\text{batch}} \subseteq \mathcal{X}_{\text{pool}}$ , we define

$$\Delta_d(\mathcal{X}_{\text{batch}}) := \max_{\mathbf{x} \in \mathcal{X}_{\text{pool}}} \min_{\mathbf{x}' \in \mathcal{X}_{\text{mode}} \cup \mathcal{X}_{\text{batch}}} d(\mathbf{x}, \mathbf{x}'),$$

$$\Delta_d^{N_{\text{batch}}} := \min_{\mathcal{X}_{\text{batch}} \subseteq \mathcal{X}_{\text{pool}}, |\mathcal{X}_{\text{batch}}| = N_{\text{batch}}} \Delta_d(\mathcal{X}_{\text{batch}}).$$

The minmax radius clustering problem is defined as finding a batch  $\mathcal{X}_{\text{batch}} \subseteq \mathcal{X}_{\text{pool}}$  such that  $\Delta_d(\mathcal{X}_{\text{batch}})$  is close to  $\Delta_d^{N_{\text{batch}}}$ . The following lemma asserts that MAXDIST yields a 2-approximation to this problem, which is in general (close to) the best possible approximation ratio for any polynomial-time algorithm unless  $\text{P} = \text{NP}$  (Feder and Greene, 1988).

---

**Algorithm D.10** MAXDIST pseudocode implementation using Algorithm D.1.
 

---

```

function INIT
     $\mathbf{c} \leftarrow (k(\mathbf{x}, \mathbf{x}))_{\mathbf{x} \in \mathcal{X}_{\text{cand}}}$ 
     $\mathbf{d} \leftarrow (\infty)_{\mathbf{x} \in \mathcal{X}_{\text{pool}}}$  ▷ Minimum squared distances
end function

function ADD( $\mathbf{x}$ )
     $\tilde{\mathbf{d}} \leftarrow (\mathbf{c}[\mathbf{x}] + \mathbf{c}[\tilde{\mathbf{x}}] - 2k(\mathbf{x}, \tilde{\mathbf{x}}))_{\tilde{\mathbf{x}} \in \mathcal{X}_{\text{pool}}}$  ▷ Compute squared kernel distances  $d_k(\mathbf{x}, \tilde{\mathbf{x}})^2$ 
     $\mathbf{d} \leftarrow \min(\mathbf{d}, \tilde{\mathbf{d}})$  ▷ element-wise minimum
end function

function NEXT
    if no point has been added yet then
        return  $\operatorname{argmax}_{\mathbf{x} \in \mathcal{X}_{\text{pool}} \setminus \mathcal{X}_{\text{batch}}} \mathbf{c}[\mathbf{x}]$ 
    end if
    return  $\operatorname{argmax}_{\mathbf{x} \in \mathcal{X}_{\text{pool}} \setminus \mathcal{X}_{\text{batch}}} \mathbf{d}[\mathbf{x}]$ 
end function
    
```

---

**Lemma D.5.** *Let  $\mathcal{X}_{\text{batch}}$  be the batch selected by MAXDIST applied to  $k$ . Then,*

$$\Delta_{d_k}(\mathcal{X}_{\text{batch}}) \leq 2\Delta_{d_k}^{N_{\text{batch}}}.$$

**Proof** For  $\mathcal{X}_{\text{mode}} = \emptyset$ , this has been proven for example in Bern and Eppstein (1996). Sener and Savarese (2018) mentioned the result for general  $\mathcal{X}_{\text{mode}}$  but it is unclear where this is proven. Therefore, we give a proof sketch here.

Let  $d := d_k$ . Let  $D$  be the distance of the last selected point in  $\mathcal{X}_{\text{batch}}$  to the remaining points in  $\mathcal{X}_{\text{batch}} \cup \mathcal{X}_{\text{mode}}$ . Then,  $\Delta_d(\mathcal{X}_{\text{batch}}) \leq D$ , because otherwise another point with a larger distance would have been chosen instead. At the same time, all points in  $\mathcal{X}_{\text{batch}}$  are at least a distance of  $D$  apart from any other point in  $\mathcal{X}_{\text{batch}} \cup \mathcal{X}_{\text{mode}}$ , since otherwise the last point would have been chosen already in an earlier step.

Now, consider a set  $\tilde{\mathcal{X}}_{\text{batch}} \subseteq \mathcal{X}_{\text{pool}}$  with  $|\tilde{\mathcal{X}}_{\text{batch}}| = N_{\text{batch}}$  such that  $\Delta_d(\tilde{\mathcal{X}}_{\text{batch}}) = \Delta_d^{N_{\text{batch}}}$ . To derive a contradiction, assume  $\Delta_d^{N_{\text{batch}}} < \Delta_d(\mathcal{X}_{\text{batch}})/2$ . Then, for every  $\mathbf{x} \in \mathcal{X}_{\text{batch}}$ , there must be  $\tilde{\mathbf{x}} \in \mathcal{X}_{\text{mode}} \cup \tilde{\mathcal{X}}_{\text{batch}}$  such that  $d(\mathbf{x}, \tilde{\mathbf{x}}) < \Delta_d(\mathcal{X}_{\text{batch}})/2$ . By our previous considerations,  $\tilde{\mathbf{x}}$  cannot be in  $\mathcal{X}_{\text{mode}}$ , so it must be in  $\tilde{\mathcal{X}}_{\text{batch}}$ . Moreover, because points in  $\mathcal{X}_{\text{batch}}$  are at least  $D$  apart, no two of them can be closer than  $\Delta_d(\mathcal{X}_{\text{batch}})/2$  to the same point in  $\tilde{\mathcal{X}}_{\text{batch}}$ , hence by the pigeonhole principle every point in  $\tilde{\mathcal{X}}_{\text{batch}}$  must have a point in  $\mathcal{X}_{\text{batch}}$  that is closer to it than  $\Delta_d(\mathcal{X}_{\text{batch}})/2$ . Now, let  $\mathbf{x}'$  be an arbitrary point in  $\mathcal{X}_{\text{pool}}$ . Then, there is  $\tilde{\mathbf{x}} \in \mathcal{X}_{\text{mode}} \cup \tilde{\mathcal{X}}_{\text{batch}}$  that is closer than  $\Delta_d(\mathcal{X}_{\text{batch}})/2$  to  $\mathbf{x}'$ . Moreover, there is  $\mathbf{x} \in \mathcal{X}_{\text{mode}} \cup \mathcal{X}_{\text{batch}}$  that is closer than  $\Delta_d(\mathcal{X}_{\text{batch}})/2$  to  $\tilde{\mathbf{x}}$ . Hence, the triangle inequality yields  $d(\mathbf{x}', \mathbf{x}) < \Delta_d(\mathcal{X}_{\text{batch}})$ , and since  $\mathbf{x}'$  was arbitrary, this is a contradiction.  $\blacksquare$

The following simple result will be helpful to prove an approximation guarantee when using sketching:

**Lemma D.6.** Let  $\mathcal{X}_{\text{pool}}$  be a finite set and let  $\alpha > 0$ . Moreover, let  $d_1, d_2$  be pseudometrics on a set  $\mathcal{X}$  such that  $d_1(\mathbf{x}, \tilde{\mathbf{x}}) \leq \alpha d_2(\mathbf{x}, \tilde{\mathbf{x}})$  for all  $\mathbf{x}, \tilde{\mathbf{x}} \in \mathcal{X}_{\text{pool}}$ . Then,

$$\begin{aligned} \forall \mathcal{X}_{\text{batch}} \subseteq \mathcal{X}_{\text{pool}} : \Delta_{d_1}(\mathcal{X}_{\text{batch}}) &\leq \alpha \Delta_{d_2}(\mathcal{X}_{\text{batch}}) , \\ \forall N_{\text{batch}} \in \{1, \dots, |\mathcal{X}_{\text{pool}}|\} : \Delta_{d_1}^{N_{\text{batch}}} &\leq \alpha \Delta_{d_2}^{N_{\text{batch}}} . \end{aligned}$$

**Proof** Let  $\mathcal{X}_{\text{batch}} \subseteq \mathcal{X}_{\text{pool}}$  and let  $\mathbf{x} \in \mathcal{X}_{\text{pool}}$ . For the element  $\mathbf{x}'' \in \mathcal{X}_{\text{mode}} \cup \mathcal{X}_{\text{batch}}$  minimizing  $d_2(\mathbf{x}, \mathbf{x}'')$ , we have

$$\min_{\mathbf{x}' \in \mathcal{X}_{\text{mode}} \cup \mathcal{X}_{\text{batch}}} d_1(\mathbf{x}, \mathbf{x}') \leq d_1(\mathbf{x}, \mathbf{x}'') \leq \alpha d_2(\mathbf{x}, \mathbf{x}'') = \alpha \min_{\mathbf{x}' \in \mathcal{X}_{\text{train}} \cup \mathcal{X}_{\text{batch}}} d_2(\mathbf{x}, \mathbf{x}') . \quad (32)$$

Applying an analogous argument to  $\mathbf{x}$  shows that  $\Delta_{d_1}(\mathcal{X}_{\text{batch}}) \leq \alpha \Delta_{d_2}(\mathcal{X}_{\text{batch}})$ . Another application to  $\mathcal{X}_{\text{batch}}$  then shows that  $\Delta_{d_1}^{N_{\text{batch}}} \leq \alpha \Delta_{d_2}^{N_{\text{batch}}}$ . ■

Finally, we obtain the following approximation guarantee with sketching:

**Theorem D.7.** Suppose that Eq. (15) holds. Then, the batch  $\mathcal{X}_{\text{batch}}$  with size  $N_{\text{batch}}$  computed by MAXDIST applied to  $k_{\rightarrow \text{sketch}(p)}$  satisfies

$$\Delta_{d_k}(\mathcal{X}_{\text{batch}}) \leq 2 \frac{1 + \varepsilon}{1 - \varepsilon} \Delta_{d_k}^{N_{\text{batch}}} .$$

**Proof** Let  $d_1 := d_k$  and  $d_2 := d_{k_{\rightarrow \text{sketch}(p)}}$ . Then,

$$\begin{aligned} \Delta_{d_1}(\mathcal{X}_{\text{batch}}) &\stackrel{\text{Lemma D.6}}{\leq} \frac{1}{1 - \varepsilon} \Delta_{d_2}(\mathcal{X}_{\text{batch}}) \stackrel{\text{Lemma D.5}}{\leq} 2 \frac{1}{1 - \varepsilon} \Delta_{d_2}^{N_{\text{batch}}} \\ &\stackrel{\text{Lemma D.6}}{\leq} 2 \frac{1 + \varepsilon}{1 - \varepsilon} \Delta_{d_1}^{N_{\text{batch}}} . \end{aligned} \quad \blacksquare$$

## D.8 KMEANSPP

Algorithm D.11 shows pseudocode for KMEANSPP. Like for MAXDIST, we obtain a runtime complexity of  $\mathcal{O}(N_{\text{pool}} N_{\text{sel}} (T_k + 1))$  and a memory complexity of  $\mathcal{O}(N_{\text{pool}})$ .

## D.9 LCMD

A pseudocode implementation of our newly proposed LCMD method is shown in Algorithm D.12. Here, the ADD method has a runtime complexity of  $\mathcal{O}(N_{\text{pool}} (T_k + 1))$  and the NEXT method has a runtime complexity of  $\mathcal{O}(N_{\text{pool}} + N_{\text{sel}})$ . The overall runtime complexity is therefore  $\mathcal{O}(N_{\text{pool}} N_{\text{sel}} (T_k + 1) + N_{\text{batch}} (N_{\text{pool}} + N_{\text{sel}})) = \mathcal{O}(N_{\text{pool}} N_{\text{sel}} (T_k + 1))$ . The memory complexity of Algorithm D.12 is  $\mathcal{O}(N_{\text{cand}})$ .

## Appendix E. Details on Experiments

In the following, we provide a more detailed description of our experimental setup and our results. All NN computations were performed with 32-bit floating-point precision, but we switched to 64-bit floating-point precision whenever posterior transformations (computations involving  $\sigma^2$ ) were involved. All MAXDET computations used the kernel-space implementation and all FRANKWOLFE computations used the feature-space implementation. All experiments were run on a workstation with four NVIDIA RTX 3090 GPUs and an AMD Ryzen Threadripper PRO 3975WX CPU with 256 GB RAM.

---

**Algorithm D.11** KMEANSPP pseudocode implementation using Algorithm D.1.
 

---

```

function INIT
     $\mathbf{c} \leftarrow (k(\mathbf{x}, \mathbf{x}))_{\mathbf{x} \in \mathcal{X}_{\text{cand}}}$ 
     $\mathbf{d} \leftarrow (\infty)_{\mathbf{x} \in \mathcal{X}_{\text{pool}}}$  ▷ Minimum squared distances
end function

function ADD( $\mathbf{x}$ )
     $\tilde{\mathbf{d}} \leftarrow (\mathbf{c}[\mathbf{x}] + \mathbf{c}[\tilde{\mathbf{x}}] - 2k(\mathbf{x}, \tilde{\mathbf{x}}))_{\tilde{\mathbf{x}} \in \mathcal{X}_{\text{pool}}}$  ▷ Compute squared kernel distances  $d_k(\mathbf{x}, \tilde{\mathbf{x}})^2$ 
     $\mathbf{d} \leftarrow \min(\mathbf{d}, \tilde{\mathbf{d}})$  ▷ element-wise minimum
end function

function NEXT
    if no point has been added yet then
        return uniform random sample from  $\mathcal{X}_{\text{pool}} \setminus \mathcal{X}_{\text{batch}}$ 
    end if
    return sample  $\mathbf{x} \in \mathcal{X}_{\text{pool}} \setminus \mathcal{X}_{\text{batch}}$  with probability proportional to  $\mathbf{d}[\mathbf{x}]$ 
end function
    
```

---

**Algorithm D.12** LCMD pseudocode implementation using Algorithm D.1.
 

---

```

function INIT
     $\mathbf{c} \leftarrow (k(\mathbf{x}, \mathbf{x}))_{\mathbf{x} \in \mathcal{X}_{\text{cand}}}$ 
     $\mathbf{d} \leftarrow (\infty)_{\mathbf{x} \in \mathcal{X}_{\text{pool}}}$  ▷ Minimum squared distances
     $\mathbf{v} \leftarrow (0)_{\mathbf{x} \in \mathcal{X}_{\text{pool}}}$  ▷ Associated cluster centers; dummy initialization
end function

function ADD( $\mathbf{x}$ )
     $\tilde{\mathbf{d}} \leftarrow (\mathbf{c}[\mathbf{x}] + \mathbf{c}[\tilde{\mathbf{x}}] - 2k(\mathbf{x}, \tilde{\mathbf{x}}))_{\tilde{\mathbf{x}} \in \mathcal{X}_{\text{pool}}}$  ▷ Compute squared kernel distances  $d_k(\mathbf{x}, \tilde{\mathbf{x}})^2$ 
     $\mathbf{v} \leftarrow (\mathbf{v}[\tilde{\mathbf{x}}] \text{ if } \mathbf{d}[\tilde{\mathbf{x}}] \leq \tilde{\mathbf{d}}[\tilde{\mathbf{x}}] \text{ else } \mathbf{x})_{\tilde{\mathbf{x}} \in \mathcal{X}_{\text{pool}}}$  ▷ Update associated cluster centers
     $\mathbf{d} \leftarrow \min(\mathbf{d}, \tilde{\mathbf{d}})$  ▷ element-wise minimum
end function

function NEXT
    if no point has been added yet then
        return  $\operatorname{argmax}_{\mathbf{x} \in \mathcal{X}_{\text{pool}} \setminus \mathcal{X}_{\text{batch}}} \mathbf{c}[\mathbf{x}]$ 
    end if
     $\mathbf{s} \leftarrow (\sum_{\tilde{\mathbf{x}} \in \mathcal{X}_{\text{pool}} \setminus \mathcal{X}_{\text{batch}}: \mathbf{v}[\tilde{\mathbf{x}}] = \mathbf{x}} \mathbf{d}[\tilde{\mathbf{x}}])_{\mathbf{x} \in \mathcal{X}_{\text{mode}} \cup \mathcal{X}_{\text{batch}}}$  ▷ Compute cluster sizes
     $s_{\text{max}} = \max_{\mathbf{x} \in \mathcal{X}_{\text{mode}} \cup \mathcal{X}_{\text{batch}}} \mathbf{s}[\mathbf{x}]$  ▷ Maximum cluster size
    return  $\operatorname{argmax}_{\mathbf{x} \in \mathcal{X}_{\text{pool}} \setminus \mathcal{X}_{\text{batch}}: \mathbf{s}[\mathbf{x}] = s_{\text{max}}} \mathbf{d}[\mathbf{x}]$ 
end function
    
```

---

## E.1 Data Sets

We selected 15 tabular regression data sets from different sources, roughly using the following criteria:

- (a) The data set should be sufficiently large after removing rows with missing values (at least 40000 samples).
- (b) The data set should be in a format suitable to perform regression, e.g., not consist of a few long time series.
- (c) The data set should not have too many categorical or text columns with many categories.
- (d) The test RMSE for randomly sampled training sets should drop substantially when going from  $N_{\text{train}} = 256$  to  $N_{\text{train}} = 17 \cdot 256$ . In our case, the selected 15 data sets differ from the other tested data sets in that the RMSE dropped at least by 14%. With this criterion, we want to exclude data sets that would not significantly affect the benchmark results, e.g., because they are too easy to learn or because they are too noisy.

An overview of the selected data sets can be found in Table E.1 and Table E.2. Our main data sources are the UCI and OpenML repositories (Dua and Graff, 2017; Vanschoren et al., 2013). The `sgemm` and `ct_slices` data sets have also been used by Tsymbalov et al. (2018). In contrast to Tsymbalov et al. (2018), we use the undirected and not the directed version of the `kegg` data set since it contains more samples and the RMSE drops more strongly between  $N_{\text{train}} = 256$  and  $N_{\text{train}} = 17 \cdot 256$ . Concerning the other four data sets used in Tsymbalov et al. (2018), we omitted the `BlogFeedback`, `YearPredictionMSD`, and `Online News Popularity` data sets due to criterion (d), and could not find the `Rosenbrock 2000D` data set online. Although the `poker` data set is originally a multi-class classification data set, we include it since it is noise-free and sufficiently difficult to learn.

Our accompanying code allows automatically downloading and processing all data sets.

## E.2 Preprocessing

We preprocess the data sets in the following way: On some data sets, we remove unwanted columns such as identifier columns, see the details in our code. We then remove rows with NaN (missing) values. Next, if necessary, we randomly subsample the data set such that it contains at most 500000 samples. We use 80% of the data set, but at most 200000 samples, for training, validation, and pool data. Of those, we initially use  $N_{\text{train}} = 256$  and  $N_{\text{valid}} = 1024$  and reserve the rest for the pool set. Subsequently, we remove columns with only a single value. We one-hot encode small categorical columns, allowing at most 300 new continuous columns, and discard larger categorical columns. On some data sets, we transform the labels  $y$ , for example by applying a logarithm or by taking the median of multiple target values, we refer to our code for further details. We standardize the labels  $y$  such that they have mean 0 and variance 1.<sup>7</sup> We preprocess the inputs  $\mathbf{x} \in \mathbb{R}^d$  as

$$x_j^{\text{processed}} := 5 \tanh\left(\frac{1}{5} \cdot \frac{x_j - \hat{\mu}_j}{\hat{\sigma}_j}\right),$$

where we compute

$$\hat{\mu}_j := \frac{1}{N_{\text{train}} + N_{\text{pool}}} \sum_{\mathbf{x} \in \mathcal{X}_{\text{train}} \cup \mathcal{X}_{\text{pool}}} x_j$$

---

7. This only leaks a negligible amount of information from the test set and in turn allows us to better compare the errors across data sets.

Short name	Initial pool set size	Test set size	Number of features
sgemm	192000	48320	14
wec_sydney	56320	14400	48
ct_slices	41520	10700	379
kegg_undir	50407	12921	27
online_video	53748	13756	26
query	158720	40000	4
poker	198720	300000	95
road	198720	234874	2
mlr_knn_rng	88123	22350	132
fried	31335	8153	10
diamonds	41872	10788	29
methane	198720	300000	33
stock	45960	11809	9
protein	35304	9146	9
sarcos	34308	8896	21

Table E.1: Data set characteristics.

$$\hat{\sigma}_j^2 := \frac{1}{N_{\text{train}} + N_{\text{pool}}} \sum_{\mathbf{x} \in \mathcal{X}_{\text{train}} \cup \mathcal{X}_{\text{pool}}} (x_j - \hat{\mu}_j)^2.$$

The motivation for the tanh function is to reduce the impact of outliers by soft-clipping the coordinates to the interval  $(-5, 5)$ .

For the sarcos data set, we only used the training data since the test data on the GPML web page (see Table E.2) is already contained in the training data. For the poker data set, we only used the test data since it contains roughly a million samples, while the training set contains around 25000 samples.

The resulting characteristics of the processed data sets can be found in Table E.1. The full names, links, and citations are contained in Table E.2.

### E.3 Neural Network Configuration

We use a fully-connected NN with two hidden layers with 512 neurons each ( $L = 3$ ,  $d_1 = d_2 = 512$ ). We employ the neural tangent parametrization as discussed in Section 2.1 with the ReLU activation function. We initialize biases to zero and weights i.i.d. from  $\mathcal{N}(0, 1)$ . For optimization, we use the Adam (Kingma and Ba, 2015) optimizer with its default parameters  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ , and let the learning rate (see below) decay linearly to zero over training. We use a mini-batch size of 256 and train for 256 epochs. After each epoch, we measure the validation RMSE on a validation set with 1024 samples. After training, we set the trained model parameters  $\boldsymbol{\theta}_T$  to the parameters from the end of the epoch where the lowest validation RMSE was attained. While the use of a large validation set might not be realistic for many data-scarce BMAL scenarios, we see this as a simple proxy for more complicated cross-validation or refitting strategies.

While the reasoning of forward variance preservation as in the well-known Kaiming initialization (He et al., 2015) suggests to set  $\sigma_w = \sqrt{2}$ , we find that smaller values of  $\sigma_w$

Short name	Source	OpenML ID	Full name	Citation
sgemm	UCI		SGEMM GPU kernel performance	Ballester-Ripoll et al. (2019)
wec_sydney	UCI		Wave Energy Converters	Neshat et al. (2018)
ct_slices	UCI		Relative location of CT slices on axial axis	Graf et al. (2011)
kegg_undir	UCI		KEGG Metabolic Reaction Network (Undirected)	Shannon et al. (2003)
online_video	UCI		Online Video Characteristics and Transcoding Time	Deneke et al. (2014)
query	UCI		Query Analytics Workloads	Anagnostopoulos et al. (2018), Savva et al. (2018)
poker	UCI		Poker Hand	—
road	UCI		3D Road Network (North Jutland, Denmark)	Kaul et al. (2013)
mlr_knn_rng	OpenML	42454	mlr_knn_rng	—
fried	OpenML	564	fried	Friedman (1991)
diamonds	OpenML	42225	diamonds	—
methane	OpenML	42701	Methane	Ślęzak et al. (2018)
stock	OpenML	1200	BNG(stock)	—
protein	OpenML	42903	physicochemical-protein	—
sarcos	GPML		SARCOS data	Vijayakumar and Schaal (2000)

Table E.2: Overview of used data sets. The second column entries are hyperlinks to the respective web pages.

can substantially improve the RMSE of the trained models. Possible explanations for this phenomenon might be that large  $\sigma_w$  increases the scale of the disturbance by the random initial function of the network (Nonnenmacher et al., 2021) or brings the NN more towards a “lazy training” regime (Chizat et al., 2019). Therefore, we manually tuned  $\sigma_w, \sigma_b$  and the initial learning rate to optimize the mean log RMSE across all data sets for RANDOM selection. We arrived at  $\sigma_w = \sigma_b = 0.2$  and an initial learning rate of 0.375.

To assess whether our insights apply to other NN configurations, we also run experiments for a fully-connected NN with the SiLU (a.k.a. Swish) activation function (Elfwing et al., 2018). Again, we use optimized hyperparameters for SiLU, specifically an initial learning rate of 0.15 as well as  $\sigma_w = 0.5, \sigma_b = 1.0$ .

### E.4 Results

Table E.5 shows averaged logarithmic error metrics and runtimes for a wide variety of configurations. Some conclusions from these results are discussed in Section 6. Table E.6 shows analogous results for our NN configuration with the SiLU instead of the ReLU activation function. Note that the results for  $k_{\text{nngp}}$  still use the NNGP for ReLU, though, since we do not know of an analytic expression of the NNGP for SiLU. Results on individual data sets for selected methods are shown in Table E.3 and Table E.4.

In this section, we also provide more plots complementing the figures from the main part of the paper. Figure E.1 shows batch size plots on individual data sets for RMSE and Figure E.2 shows learning curve plots on individual data sets for the 99% quantile. Moreover, Figure E.3 and Figure E.4 allow comparing two methods across data sets on RMSE and MAXE, respectively.

The estimated standard deviations of the mean estimators in Figure 1, Figure 3, Figure 4, Figure 5, Figure E.1 and Figure E.2 are computed as follows: Consider random variables  $X_{ij}$  representing the log metric values on repetition  $i$  and data set  $j$ . Then, it is well-known that for the mean estimator  $\hat{\mu}_j := \frac{1}{20} \sum_{i=1}^{20} X_{ij}$ , an unbiased estimator of its variance is given by

$$\hat{\sigma}_j^2 := \frac{1}{20-1} \sum_{i=1}^{20} (X_{ij} - \hat{\mu}_j)^2$$

Since all mean estimators  $\hat{\mu}_j$  are independent, the variance of the total mean estimator  $\hat{\mu} := \frac{1}{15} \sum_{j=1}^{15} \hat{\mu}_j$  can be estimated as

$$\hat{\sigma}^2 := \frac{1}{15^2} \sum_{j=1}^{15} \hat{\sigma}_j^2.$$

Our plots hence show  $\hat{\sigma}$  as the estimated standard deviation of the mean estimator  $\hat{\mu}$ .



Data set	RANDOM	MAXDIAG	MAXDET-P	BAIT-F-P	FRANKWOLFE-P	MAXDIST-P	KMEANSPP-P	LCMD-TP (ours)
ct_slices	0.141	0.123	0.085	0.076	0.088	0.085	0.081	<b>0.072</b>
diamonds	0.173	0.169	0.166	<b>0.161</b>	0.162	0.166	0.162	<b>0.161</b>
fried	0.230	0.231	0.228	<b>0.227</b>	0.228	0.229	0.229	0.228
kegg_undir	0.380	0.346	0.245	0.222	0.248	0.243	0.227	<b>0.220</b>
methane	0.733	0.770	0.736	0.714	0.714	0.740	0.713	<b>0.708</b>
mlr_knn_rng	0.294	0.326	0.211	0.183	0.199	0.209	0.190	<b>0.176</b>
online_video	0.263	0.190	0.159	<b>0.149</b>	0.159	0.158	0.153	0.152
poker	0.806	0.803	<b>0.742</b>	0.751	0.797	0.754	0.794	0.797
protein	0.763	0.793	0.782	<b>0.757</b>	0.761	0.782	<b>0.757</b>	0.759
query	0.058	0.082	0.066	0.057	0.060	0.066	0.057	<b>0.053</b>
road	<b>0.586</b>	0.702	0.625	0.592	0.606	0.624	0.598	0.591
sarcos	0.181	0.190	0.176	0.164	0.168	0.176	0.167	<b>0.163</b>
sgemm	0.152	0.185	0.155	0.142	0.144	0.153	0.144	<b>0.140</b>
stock	0.531	0.541	0.540	0.529	0.527	0.540	<b>0.526</b>	0.528
wec_sydney	0.027	0.034	0.030	<b>0.025</b>	0.027	0.030	0.026	0.028

Table E.3: This table shows the averaged (non-logarithmic) RMSEs per data set, averaged over all repetitions, and BMAL steps, for each of the selection methods with kernels as in Table 6.

Data set	RANDOM	MAXDIAG	MAXDET-P	BAIT-F-P	FRANKWOLFE-P	MAXDIST-P	KMEANSPP-P	LCMD-TP (ours)
ct_slices	0.093	0.069	0.053	0.044	0.050	0.052	0.046	<b>0.038</b>
diamonds	0.166	0.155	0.153	<b>0.152</b>	<b>0.152</b>	0.153	0.153	<b>0.152</b>
fried	0.221	0.219	<b>0.218</b>	0.219	0.219	0.219	0.220	0.219
kegg_undir	0.291	0.225	0.180	0.165	0.173	0.179	0.166	<b>0.154</b>
methane	0.692	0.731	0.704	0.675	0.670	0.709	0.669	<b>0.661</b>
mlr_knn_rng	0.208	0.168	0.126	0.112	0.119	0.131	0.108	<b>0.105</b>
online_video	0.206	0.137	0.116	0.106	0.111	0.115	0.109	<b>0.105</b>
poker	0.600	0.598	<b>0.533</b>	0.537	0.578	0.547	0.558	0.589
protein	0.727	0.758	0.751	0.721	0.729	0.750	<b>0.717</b>	0.721
query	0.040	0.061	0.054	0.042	0.046	0.052	0.040	<b>0.037</b>
road	<b>0.538</b>	0.671	0.582	0.544	0.571	0.570	0.549	0.551
sarcos	0.155	0.162	0.155	0.140	0.143	0.154	0.142	<b>0.138</b>
sgemm	0.100	0.126	0.107	0.098	0.097	0.106	0.096	<b>0.092</b>
stock	0.511	0.517	0.518	0.508	<b>0.506</b>	0.519	<b>0.506</b>	<b>0.506</b>
wec_sydney	0.021	0.023	0.021	<b>0.020</b>	0.021	0.021	<b>0.020</b>	<b>0.020</b>

Table E.4: This table shows the (non-logarithmic) RMSEs after the last BMAL step per data set, averaged over all repetitions, for each of the selection methods with kernels as in Table 6.

Selection method	Kernel	MAE	RMSE	95%	99%	MAXE	avg. time [s]
RANDOM	—	-1.934	-1.401	-0.766	-0.163	1.107	0.001
MAXDIAC	$k_{\text{grad}} \rightarrow \text{sketch}(512) \rightarrow \text{acs-rf}(512)$	-1.777	-1.370	-0.690	-0.189	0.978	0.650
MAXDIAC	$k_{\text{grad}} \rightarrow \text{sketch}(512) \rightarrow \mathcal{X}_{\text{train}}$	-1.777	-1.369	-0.690	-0.186	0.986	0.551
MAXDIAC	$k_{\text{grad}} \rightarrow \text{ens}(3) \rightarrow \text{sketch}(512) \rightarrow \mathcal{X}_{\text{train}}$	-1.768	-1.366	-0.685	-0.188	0.970	1.392
MAXDIAC	$k_{\text{grad}} \rightarrow \mathcal{X}_{\text{train}}$	-1.766	-1.355	-0.675	-0.168	0.996	4.108
MAXDIAC	$k_{\text{grad}} \rightarrow \text{sketch}(512) \rightarrow \text{acs-grad}$	-1.751	-1.345	-0.664	-0.163	1.007	0.553
MAXDIAC	$k_{\text{grad}} \rightarrow \text{sketch}(512) \rightarrow \text{acs-rf-hyper}(512)$	-1.743	-1.334	-0.656	-0.148	1.021	0.650
MAXDIAC	$k_{\parallel} \rightarrow \text{acs-rf}(512)$	-1.713	-1.286	-0.606	-0.084	1.052	0.030
MAXDIAC	$k_{\parallel} \rightarrow \mathcal{X}_{\text{train}}$	-1.722	-1.285	-0.614	-0.077	1.080	0.142
MAXDIAC	$k_{\text{nnnp}} \rightarrow \mathcal{X}_{\text{train}}$	-1.754	-1.272	-0.606	-0.044	1.088	2.592
MAXDIAC	$k_{\text{lin}} \rightarrow \mathcal{X}_{\text{train}}$	-1.585	-1.103	-0.426	0.143	1.222	0.007
MAXDET-TP	$k_{\text{grad}} \rightarrow \text{scale}(\mathcal{X}_{\text{train}})$	-1.930	-1.522	-0.855	-0.356	0.856	8.883
MAXDET-P	$k_{\text{grad}} \rightarrow \text{sketch}(512) \rightarrow \mathcal{X}_{\text{train}}$	-1.915	-1.512	-0.844	-0.350	0.867	0.770
MAXDET-P	$k_{\parallel} \rightarrow \text{ens}(3) \rightarrow \text{sketch}(512) \rightarrow \mathcal{X}_{\text{train}}$	-1.931	-1.504	-0.849	-0.337	0.873	0.673
MAXDET-P	$k_{\text{grad}} \rightarrow \text{ens}(3) \rightarrow \text{sketch}(512) \rightarrow \mathcal{X}_{\text{train}}$	-1.895	-1.500	-0.830	-0.344	0.868	1.608
MAXDET-P	$k_{\text{grad}} \rightarrow \text{sketch}(512) \rightarrow \text{acs-rf}(512)$	-1.893	-1.492	-0.820	-0.323	0.886	0.869
MAXDET-P	$k_{\text{grad}} \rightarrow \text{sketch}(512) \rightarrow \text{acs-grad}$	-1.872	-1.475	-0.802	-0.311	0.878	0.893
MAXDET-P	$k_{\text{grad}} \rightarrow \text{sketch}(512) \rightarrow \text{acs-rf-hyper}(512)$	-1.872	-1.475	-0.803	-0.310	0.888	0.872
MAXDET-P	$k_{\parallel} \rightarrow \mathcal{X}_{\text{train}}$	-1.895	-1.463	-0.808	-0.288	0.916	0.370
MAXDET-P	$k_{\parallel} \rightarrow \text{acs-rf}(512)$	-1.876	-1.443	-0.792	-0.264	0.961	0.518
MAXDET-TP	$k_{\text{nnnp}} \rightarrow \text{scale}(\mathcal{X}_{\text{train}})$	-1.848	-1.358	-0.702	-0.133	1.028	5.449
MAXDET-P	$k_{\text{lin}} \rightarrow \mathcal{X}_{\text{train}}$	-1.682	-1.187	-0.524	0.055	1.191	0.170
BAIT-F-P	$k_{\text{grad}} \rightarrow \text{ens}(3) \rightarrow \text{sketch}(512) \rightarrow \mathcal{X}_{\text{train}}$	-2.011	-1.587	-0.927	<b>-0.419</b>	0.859	2.346
BAIT-F-P	$k_{\text{grad}} \rightarrow \text{sketch}(512) \rightarrow \mathcal{X}_{\text{train}}$	-2.013	-1.585	-0.926	-0.412	0.862	1.508
BAIT-FB-P	$k_{\text{grad}} \rightarrow \text{sketch}(512) \rightarrow \mathcal{X}_{\text{train}}$	-2.007	-1.584	-0.921	-0.411	<b>0.852</b>	3.050
BAIT-F-P	$k_{\parallel} \rightarrow \text{ens}(3) \rightarrow \text{sketch}(512) \rightarrow \mathcal{X}_{\text{train}}$	<b>-2.041</b>	-1.583	-0.940	-0.400	0.880	1.408
BAIT-F-P	$k_{\parallel} \rightarrow \mathcal{X}_{\text{train}}$	-2.003	-1.545	-0.900	-0.362	0.891	1.149
BAIT-FB-P	$k_{\parallel} \rightarrow \mathcal{X}_{\text{train}}$	-1.998	-1.541	-0.895	-0.357	0.888	2.731
BAIT-F-P	$k_{\text{lin}} \rightarrow \mathcal{X}_{\text{train}}$	-1.721	-1.220	-0.562	0.022	1.162	0.232
FRANKWOLFE-P	$k_{\text{grad}} \rightarrow \text{sketch}(512) \rightarrow \text{acs-rf-hyper}(512)$	-1.977	-1.542	-0.892	-0.362	0.918	0.823
FRANKWOLFE-P	$k_{\text{grad}} \rightarrow \text{sketch}(512) \rightarrow \text{acs-grad} \rightarrow \text{sketch}(512)$	-1.999	-1.520	-0.879	-0.317	1.015	0.914
FRANKWOLFE-P	$k_{\parallel} \rightarrow \text{acs-rf}(512)$	-1.992	-1.519	-0.883	-0.321	1.022	0.421
FRANKWOLFE-P	$k_{\text{grad}} \rightarrow \text{sketch}(512) \rightarrow \text{acs-rf}(512)$	-1.995	-1.499	-0.864	-0.287	1.055	0.825
FRANKWOLFE-P	$k_{\parallel} \rightarrow \text{acs-grad} \rightarrow \text{sketch}(512)$	-1.943	-1.446	-0.807	-0.226	1.085	0.517
FRANKWOLFE-P	$k_{\parallel} \rightarrow \text{acs-rf-hyper}(512)$	-1.937	-1.439	-0.793	-0.225	1.016	0.421
FRANKWOLFE-P	$k_{\text{grad}} \rightarrow \text{sketch}(512) \rightarrow \mathcal{X}_{\text{train}}$	-1.924	-1.410	-0.765	-0.178	1.134	0.723
FRANKWOLFE-P	$k_{\parallel} \rightarrow \mathcal{X}_{\text{train}}$	-1.896	-1.391	-0.752	-0.158	1.139	0.326
MAXDIST-TP	$k_{\parallel} \rightarrow \text{ens}(3) \rightarrow \text{sketch}(512)$	-1.948	-1.518	-0.860	-0.348	0.905	0.655
MAXDIST-P	$k_{\text{grad}} \rightarrow \text{sketch}(512) \rightarrow \mathcal{X}_{\text{train}}$	-1.916	-1.514	-0.845	-0.351	0.866	0.713
MAXDIST-TP	$k_{\text{grad}} \rightarrow \text{sketch}(512)$	-1.899	-1.506	-0.838	-0.347	0.868	0.653
MAXDIST-TP	$k_{\text{grad}}$	-1.894	-1.503	-0.834	-0.342	0.866	2.347
MAXDIST-TP	$k_{\text{grad}} \rightarrow \text{ens}(3) \rightarrow \text{sketch}(512)$	-1.889	-1.498	-0.831	-0.342	0.871	0.743
MAXDIST-TP	$k_{\parallel}$	-1.924	-1.491	-0.832	-0.307	0.927	0.621
MAXDIST-P	$k_{\text{grad}} \rightarrow \text{sketch}(512) \rightarrow \text{acs-rf}(512)$	-1.893	-1.491	-0.819	-0.322	0.891	0.810
MAXDIST-P	$k_{\text{grad}} \rightarrow \text{sketch}(512) \rightarrow \text{acs-grad}$	-1.873	-1.477	-0.802	-0.312	0.888	0.832
MAXDIST-P	$k_{\text{grad}} \rightarrow \text{sketch}(512) \rightarrow \text{acs-rf-hyper}(512)$	-1.867	-1.472	-0.798	-0.306	0.895	0.811
MAXDIST-P	$k_{\parallel} \rightarrow \mathcal{X}_{\text{train}}$	-1.889	-1.459	-0.807	-0.283	0.947	0.309
MAXDIST-P	$k_{\parallel} \rightarrow \text{acs-rf}(512)$	-1.863	-1.430	-0.777	-0.247	0.980	0.410
MAXDIST-TP	$k_{\text{lin}}$	-1.888	-1.398	-0.749	-0.172	1.034	0.242
MAXDIST-TP	$k_{\text{nnnp}}$	-1.876	-1.386	-0.735	-0.159	1.038	1.315
KMEANSPP-TP	$k_{\text{grad}}$	-2.025	-1.569	-0.927	-0.378	0.966	2.357
KMEANSPP-P	$k_{\text{grad}} \rightarrow \text{sketch}(512) \rightarrow \text{acs-rf}(512)$	-2.006	-1.569	-0.912	-0.385	0.929	0.836
KMEANSPP-TP	$k_{\text{grad}} \rightarrow \text{ens}(3) \rightarrow \text{sketch}(512)$	-2.025	-1.569	-0.926	-0.377	0.967	0.754
KMEANSPP-TP	$k_{\text{grad}} \rightarrow \text{sketch}(512)$	-2.023	-1.567	-0.925	-0.376	0.967	0.663
KMEANSPP-P	$k_{\text{grad}} \rightarrow \text{sketch}(512) \rightarrow \text{acs-grad}$	-2.008	-1.558	-0.905	-0.366	0.979	0.859
KMEANSPP-P	$k_{\text{grad}} \rightarrow \text{sketch}(512) \rightarrow \text{acs-rf-hyper}(512)$	-1.994	-1.554	-0.899	-0.370	0.957	0.836
KMEANSPP-P	$k_{\text{grad}} \rightarrow \text{sketch}(512) \rightarrow \mathcal{X}_{\text{train}}$	-2.020	-1.549	-0.905	-0.348	0.997	0.738
KMEANSPP-P	$k_{\parallel} \rightarrow \mathcal{X}_{\text{train}}$	-2.007	-1.530	-0.895	-0.329	1.008	0.335
KMEANSPP-P	$k_{\parallel} \rightarrow \text{acs-rf}(512)$	-1.986	-1.529	-0.889	-0.339	0.977	0.435
KMEANSPP-TP	$k_{\parallel} \rightarrow \text{ens}(3) \rightarrow \text{sketch}(512)$	-2.020	-1.522	-0.890	-0.317	1.014	0.666
KMEANSPP-TP	$k_{\parallel}$	-2.015	-1.521	-0.887	-0.316	1.014	0.632
KMEANSPP-TP	$k_{\text{nnnp}}$	-1.969	-1.446	-0.817	-0.215	1.072	1.319
KMEANSPP-TP	$k_{\text{lin}}$	-1.968	-1.441	-0.816	-0.212	1.077	0.252
LCMD-TP (ours)	$k_{\text{grad}} \rightarrow \text{ens}(3) \rightarrow \text{sketch}(512)$	-2.040	<b>-1.594</b>	<b>-0.947</b>	-0.408	0.920	1.073
LCMD-TP (ours)	$k_{\text{grad}}$	-2.038	<b>-1.594</b>	-0.946	-0.408	0.908	2.714
LCMD-TP (ours)	$k_{\text{grad}} \rightarrow \text{sketch}(512)$	-2.033	-1.590	-0.941	-0.404	0.917	0.981
LCMD-TP (ours)	$k_{\parallel} \rightarrow \text{ens}(3) \rightarrow \text{sketch}(512)$	-2.026	-1.555	-0.912	-0.358	0.952	0.984
LCMD-P (ours)	$k_{\text{grad}} \rightarrow \text{sketch}(512) \rightarrow \text{acs-rf}(512)$	-1.992	-1.547	-0.905	-0.369	0.974	0.875
LCMD-TP (ours)	$k_{\parallel}$	-2.017	-1.544	-0.902	-0.346	0.961	0.948
LCMD-P (ours)	$k_{\text{grad}} \rightarrow \text{sketch}(512) \rightarrow \mathcal{X}_{\text{train}}$	-1.940	-1.534	-0.869	-0.371	0.864	0.774
LCMD-P (ours)	$k_{\parallel} \rightarrow \text{acs-rf}(512)$	-1.969	-1.513	-0.885	-0.331	1.010	0.473
LCMD-P (ours)	$k_{\text{grad}} \rightarrow \text{sketch}(512) \rightarrow \text{acs-rf-hyper}(512)$	-1.898	-1.501	-0.827	-0.336	0.878	0.876
LCMD-P (ours)	$k_{\text{grad}} \rightarrow \text{sketch}(512) \rightarrow \text{acs-grad}$	-1.896	-1.496	-0.827	-0.334	0.880	0.899
LCMD-P (ours)	$k_{\parallel} \rightarrow \mathcal{X}_{\text{train}}$	-1.920	-1.485	-0.840	-0.311	0.935	0.372
LCMD-TP (ours)	$k_{\text{nnnp}}$	-1.960	-1.447	-0.811	-0.216	1.043	1.587
LCMD-TP (ours)	$k_{\text{lin}}$	-1.958	-1.445	-0.810	-0.215	1.036	0.525

Table E.5: This table shows the performance and runtime of different combinations of selection methods and kernels. The columns labeled “MAE” to “MAXE” contain averaged logarithmic values of the corresponding metrics, averaged over all data sets, repetitions, and BMAL steps. For ensemble kernels, the metrics of the individual ensemble members were averaged to isolate the effect of ensembling on the batch selection. Runtimes were measured at one of the 20 repetitions where only one process was started per GPU, and are averaged over all BMAL steps and data sets. The employed hardware is described in Section E.

DEEP BATCH ACTIVE LEARNING FOR REGRESSION

Selection method	Kernel	MAE	RMSE	95%	99%	MAXE	avg. time [s]
RANDOM	—	-1.923	-1.406	-0.774	-0.178	1.119	0.001
MAXDIAC	$k_{\text{grad}} \rightarrow \text{ens}(3) \rightarrow \text{sketch}(512) \rightarrow \mathcal{X}_{\text{train}}$	-1.753	-1.358	-0.690	-0.205	0.956	1.393
MAXDIAC	$k_{\text{grad}} \rightarrow \text{sketch}(512) \rightarrow \mathcal{X}_{\text{train}}$	-1.751	-1.351	-0.682	-0.192	0.961	0.551
MAXDIAC	$k_{\text{grad}} \rightarrow \text{sketch}(512) \rightarrow \text{acs-rf}(512)$	-1.749	-1.350	-0.680	-0.189	0.962	0.651
MAXDIAC	$k_{\text{grad}} \rightarrow \mathcal{X}_{\text{train}}$	-1.749	-1.346	-0.677	-0.182	0.967	4.107
MAXDIAC	$k_{\text{grad}} \rightarrow \text{sketch}(512) \rightarrow \text{acs-grad}$	-1.735	-1.342	-0.671	-0.184	0.961	0.553
MAXDIAC	$k_{\text{nngp}} \rightarrow \mathcal{X}_{\text{train}}$	-1.768	-1.301	-0.634	-0.095	1.054	2.584
MAXDIAC	$k_{\text{ll}} \rightarrow \mathcal{X}_{\text{train}}$	-1.744	-1.300	-0.631	-0.100	1.037	0.142
MAXDIAC	$k_{\text{ll}} \rightarrow \text{acs-rf}(512)$	-1.713	-1.286	-0.606	-0.084	1.052	0.031
MAXDIAC	$k_{\text{grad}} \rightarrow \text{sketch}(512) \rightarrow \text{acs-rf-hyper}(512)$	-1.675	-1.279	-0.604	-0.111	1.007	0.649
MAXDIAC	$k_{\text{lin}} \rightarrow \mathcal{X}_{\text{train}}$	-1.622	-1.158	-0.483	0.068	1.156	0.007
MAXDET-TP	$k_{\text{grad}} \rightarrow \text{scale}(\mathcal{X}_{\text{train}})$	-1.955	-1.546	-0.895	-0.400	0.834	8.914
MAXDET-P	$k_{\text{grad}} \rightarrow \text{ens}(3) \rightarrow \text{sketch}(512) \rightarrow \mathcal{X}_{\text{train}}$	-1.913	-1.523	-0.867	-0.390	0.843	1.607
MAXDET-P	$k_{\text{grad}} \rightarrow \text{sketch}(512) \rightarrow \mathcal{X}_{\text{train}}$	-1.915	-1.520	-0.864	-0.381	0.846	0.770
MAXDET-P	$k_{\text{grad}} \rightarrow \text{sketch}(512) \rightarrow \text{acs-rf}(512)$	-1.910	-1.514	-0.857	-0.372	0.855	0.868
MAXDET-P	$k_{\text{grad}} \rightarrow \text{sketch}(512) \rightarrow \text{acs-grad}$	-1.886	-1.503	-0.840	-0.370	0.846	0.890
MAXDET-P	$k_{\text{ll}} \rightarrow \text{ens}(3) \rightarrow \text{sketch}(512) \rightarrow \mathcal{X}_{\text{train}}$	-1.921	-1.485	-0.823	-0.304	0.894	0.675
MAXDET-P	$k_{\text{grad}} \rightarrow \text{sketch}(512) \rightarrow \text{acs-rf-hyper}(512)$	-1.847	-1.468	-0.804	-0.335	0.863	0.871
MAXDET-P	$k_{\text{ll}} \rightarrow \mathcal{X}_{\text{train}}$	-1.910	-1.467	-0.807	-0.281	0.913	0.371
MAXDET-P	$k_{\text{ll}} \rightarrow \text{acs-rf}(512)$	-1.905	-1.459	-0.801	-0.272	0.924	0.470
MAXDET-TP	$k_{\text{nngp}} \rightarrow \text{scale}(\mathcal{X}_{\text{train}})$	-1.869	-1.401	-0.737	-0.198	0.980	5.456
MAXDET-P	$k_{\text{lin}} \rightarrow \mathcal{X}_{\text{train}}$	-1.714	-1.239	-0.574	-0.020	1.118	0.170
BAIT-F-P	$k_{\text{grad}} \rightarrow \text{ens}(3) \rightarrow \text{sketch}(512) \rightarrow \mathcal{X}_{\text{train}}$	-2.025	-1.594	-0.948	<b>-0.436</b>	0.855	2.342
BAIT-F-P	$k_{\text{grad}} \rightarrow \text{sketch}(512) \rightarrow \mathcal{X}_{\text{train}}$	-2.023	-1.588	-0.943	-0.429	0.859	1.505
BAIT-FB-P	$k_{\text{grad}} \rightarrow \text{sketch}(512) \rightarrow \mathcal{X}_{\text{train}}$	-2.019	-1.585	-0.938	-0.426	0.853	3.051
BAIT-F-P	$k_{\text{ll}} \rightarrow \text{ens}(3) \rightarrow \text{sketch}(512) \rightarrow \mathcal{X}_{\text{train}}$	-2.002	-1.541	-0.885	-0.345	0.894	1.407
BAIT-F-P	$k_{\text{ll}} \rightarrow \mathcal{X}_{\text{train}}$	-1.989	-1.524	-0.869	-0.324	0.913	1.147
BAIT-FB-P	$k_{\text{ll}} \rightarrow \mathcal{X}_{\text{train}}$	-1.989	-1.522	-0.868	-0.323	0.917	2.733
BAIT-F-P	$k_{\text{lin}} \rightarrow \mathcal{X}_{\text{train}}$	-1.742	-1.264	-0.601	-0.045	1.110	0.233
FRANKWOLFE-P	$k_{\text{grad}} \rightarrow \text{sketch}(512) \rightarrow \text{acs-rf-hyper}(512)$	-1.961	-1.529	-0.882	-0.363	0.922	0.822
FRANKWOLFE-P	$k_{\text{grad}} \rightarrow \text{sketch}(512) \rightarrow \text{acs-rf}(512)$	-1.977	-1.502	-0.864	-0.299	1.015	0.824
FRANKWOLFE-P	$k_{\text{grad}} \rightarrow \text{sketch}(512) \rightarrow \text{acs-grad} \rightarrow \text{sketch}(512)$	-1.960	-1.481	-0.840	-0.279	1.049	0.915
FRANKWOLFE-P	$k_{\text{ll}} \rightarrow \text{acs-rf}(512)$	-1.959	-1.467	-0.830	-0.255	1.031	0.422
FRANKWOLFE-P	$k_{\text{ll}} \rightarrow \text{acs-rf-hyper}(512)$	-1.923	-1.437	-0.797	-0.227	1.016	0.420
FRANKWOLFE-P	$k_{\text{ll}} \rightarrow \text{acs-grad} \rightarrow \text{sketch}(512)$	-1.911	-1.413	-0.771	-0.191	1.092	0.519
FRANKWOLFE-P	$k_{\text{grad}} \rightarrow \text{sketch}(512) \rightarrow \mathcal{X}_{\text{train}}$	-1.881	-1.384	-0.736	-0.157	1.129	0.725
FRANKWOLFE-P	$k_{\text{ll}} \rightarrow \mathcal{X}_{\text{train}}$	-1.859	-1.358	-0.711	-0.128	1.143	0.324
MAXDIST-TP	$k_{\text{grad}} \rightarrow \text{ens}(3) \rightarrow \text{sketch}(512)$	-1.919	-1.540	-0.889	-0.416	<b>0.818</b>	0.744
MAXDIST-TP	$k_{\text{grad}}$	-1.919	-1.539	-0.888	-0.413	0.823	2.351
MAXDIST-TP	$k_{\text{grad}} \rightarrow \text{sketch}(512)$	-1.918	-1.536	-0.886	-0.409	0.820	0.652
MAXDIST-TP	$k_{\text{ll}} \rightarrow \text{ens}(3) \rightarrow \text{sketch}(512)$	-1.958	-1.526	-0.879	-0.360	0.876	0.655
MAXDIST-P	$k_{\text{grad}} \rightarrow \text{sketch}(512) \rightarrow \mathcal{X}_{\text{train}}$	-1.913	-1.518	-0.861	-0.378	0.849	0.712
MAXDIST-TP	$k_{\text{ll}}$	-1.951	-1.515	-0.870	-0.346	0.888	0.621
MAXDIST-P	$k_{\text{grad}} \rightarrow \text{sketch}(512) \rightarrow \text{acs-rf}(512)$	-1.906	-1.508	-0.851	-0.364	0.865	0.811
MAXDIST-P	$k_{\text{grad}} \rightarrow \text{sketch}(512) \rightarrow \text{acs-grad}$	-1.890	-1.507	-0.844	-0.373	0.843	0.833
MAXDIST-P	$k_{\text{ll}} \rightarrow \mathcal{X}_{\text{train}}$	-1.913	-1.468	-0.810	-0.282	0.917	0.311
MAXDIST-P	$k_{\text{grad}} \rightarrow \text{sketch}(512) \rightarrow \text{acs-rf-hyper}(512)$	-1.846	-1.462	-0.802	-0.327	0.868	0.810
MAXDIST-P	$k_{\text{ll}} \rightarrow \text{acs-rf}(512)$	-1.899	-1.452	-0.794	-0.264	0.937	0.409
MAXDIST-TP	$k_{\text{ll}} \rightarrow \text{acs-rf}(512)$	-1.905	-1.435	-0.777	-0.233	0.982	0.238
MAXDIST-TP	$k_{\text{lin}}$	-1.897	-1.426	-0.768	-0.224	0.989	1.286
KMEANSPP-TP	$k_{\text{grad}} \rightarrow \text{ens}(3) \rightarrow \text{sketch}(512)$	-2.022	-1.566	-0.931	-0.382	0.971	0.755
KMEANSPP-TP	$k_{\text{grad}}$	-2.023	-1.566	-0.932	-0.381	0.969	2.360
KMEANSPP-TP	$k_{\text{grad}} \rightarrow \text{sketch}(512)$	-2.022	-1.566	-0.932	-0.381	0.969	0.662
KMEANSPP-P	$k_{\text{grad}} \rightarrow \text{sketch}(512) \rightarrow \text{acs-rf}(512)$	-2.001	-1.558	-0.914	-0.381	0.920	0.837
KMEANSPP-P	$k_{\text{grad}} \rightarrow \text{sketch}(512) \rightarrow \text{acs-grad}$	-2.006	-1.542	-0.902	-0.351	0.972	0.858
KMEANSPP-P	$k_{\text{grad}} \rightarrow \text{sketch}(512) \rightarrow \text{acs-rf-hyper}(512)$	-1.987	-1.541	-0.898	-0.363	0.972	0.836
KMEANSPP-P	$k_{\text{grad}} \rightarrow \text{sketch}(512) \rightarrow \mathcal{X}_{\text{train}}$	-2.006	-1.531	-0.895	-0.334	0.996	0.736
KMEANSPP-TP	$k_{\text{ll}} \rightarrow \text{ens}(3) \rightarrow \text{sketch}(512)$	-2.011	-1.525	-0.894	-0.325	1.000	0.665
KMEANSPP-TP	$k_{\text{ll}}$	-2.007	-1.523	-0.889	-0.320	1.010	0.632
KMEANSPP-P	$k_{\text{ll}} \rightarrow \text{acs-rf}(512)$	-1.982	-1.501	-0.861	-0.299	0.985	0.434
KMEANSPP-P	$k_{\text{ll}} \rightarrow \mathcal{X}_{\text{train}}$	-1.980	-1.484	-0.849	-0.275	1.020	0.335
KMEANSPP-TP	$k_{\text{nngp}}$	-1.968	-1.465	-0.829	-0.246	1.051	1.298
KMEANSPP-TP	$k_{\text{lin}}$	-1.966	-1.464	-0.827	-0.248	1.057	0.249
LCMD-TP (ours)	$k_{\text{grad}} \rightarrow \text{ens}(3) \rightarrow \text{sketch}(512)$	<b>-2.041</b>	<b>-1.600</b>	<b>-0.961</b>	-0.425	0.901	1.071
LCMD-TP (ours)	$k_{\text{grad}}$	-2.038	-1.598	-0.958	-0.423	0.899	2.712
LCMD-TP (ours)	$k_{\text{grad}} \rightarrow \text{sketch}(512)$	-2.038	-1.597	-0.957	-0.422	0.898	0.977
LCMD-TP (ours)	$k_{\text{ll}} \rightarrow \text{ens}(3) \rightarrow \text{sketch}(512)$	-2.027	-1.554	-0.916	-0.359	0.941	0.979
LCMD-TP (ours)	$k_{\text{ll}}$	-2.022	-1.550	-0.911	-0.357	0.944	0.946
LCMD-P (ours)	$k_{\text{grad}} \rightarrow \text{sketch}(512) \rightarrow \text{acs-rf}(512)$	-1.982	-1.532	-0.889	-0.348	0.945	0.874
LCMD-P (ours)	$k_{\text{grad}} \rightarrow \text{sketch}(512) \rightarrow \mathcal{X}_{\text{train}}$	-1.930	-1.531	-0.880	-0.392	0.847	0.774
LCMD-P (ours)	$k_{\text{grad}} \rightarrow \text{sketch}(512) \rightarrow \text{acs-grad}$	-1.907	-1.520	-0.862	-0.389	0.849	0.898
LCMD-P (ours)	$k_{\text{grad}} \rightarrow \text{sketch}(512) \rightarrow \text{acs-rf-hyper}(512)$	-1.905	-1.508	-0.856	-0.371	0.870	0.877
LCMD-TP (ours)	$k_{\text{lin}}$	-1.971	-1.484	-0.836	-0.274	0.991	0.517
LCMD-TP (ours)	$k_{\text{nngp}}$	-1.970	-1.482	-0.834	-0.271	0.996	1.560
LCMD-P (ours)	$k_{\text{ll}} \rightarrow \mathcal{X}_{\text{train}}$	-1.928	-1.481	-0.825	-0.296	0.919	0.370
LCMD-P (ours)	$k_{\text{ll}} \rightarrow \text{acs-rf}(512)$	-1.963	-1.476	-0.840	-0.272	0.986	0.475

Table E.6: This table shows the performance and runtime of different combinations of selection methods and kernels for the SiLU activation function. The columns labeled “MAE” to “MAXE” contain averaged logarithmic values of the corresponding metrics, averaged over all data sets, repetitions and BMAL steps. For ensembled kernels, the metrics of the individual ensemble members were averaged to isolate the effect of ensembling on the batch selection. Runtimes were measured at one of the 20 repetitions where only one process was started per GPU, and are averaged over all BMAL steps and data sets. The employed hardware is described in Section E.

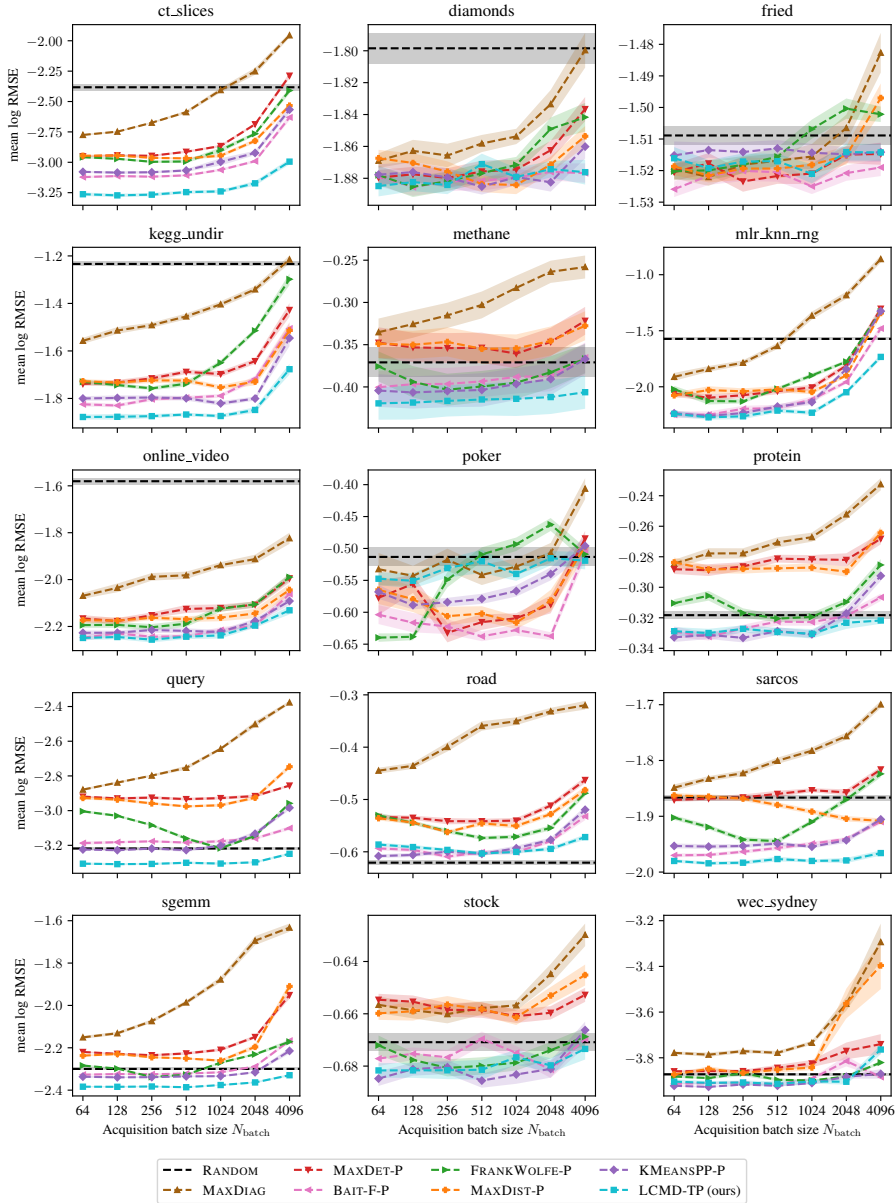


Figure E.1: This figure shows how much the final accuracy of different BMDAL methods deteriorates on individual data sets when fewer BMAL steps with larger batch sizes are used. Specifically, we use different selection methods with the corresponding kernels from Table 6, starting with  $N_{\text{train}} = 256$  and then performing  $2^m$  BMAL steps with batch size  $N_{\text{batch}} = 2^{12-m}$  for  $m \in \{0, \dots, 6\}$ , such that the final training set size is 4352 in each case. The plots show the final logarithmic error metric, averaged over all repetitions. Note that the performance of RANDOM selection does not depend on  $N_{\text{batch}}$  but only on the final training set size, hence it is shown as a constant line here. The shaded area corresponds to one estimated standard deviation of the mean estimator, cf. Section E.4.

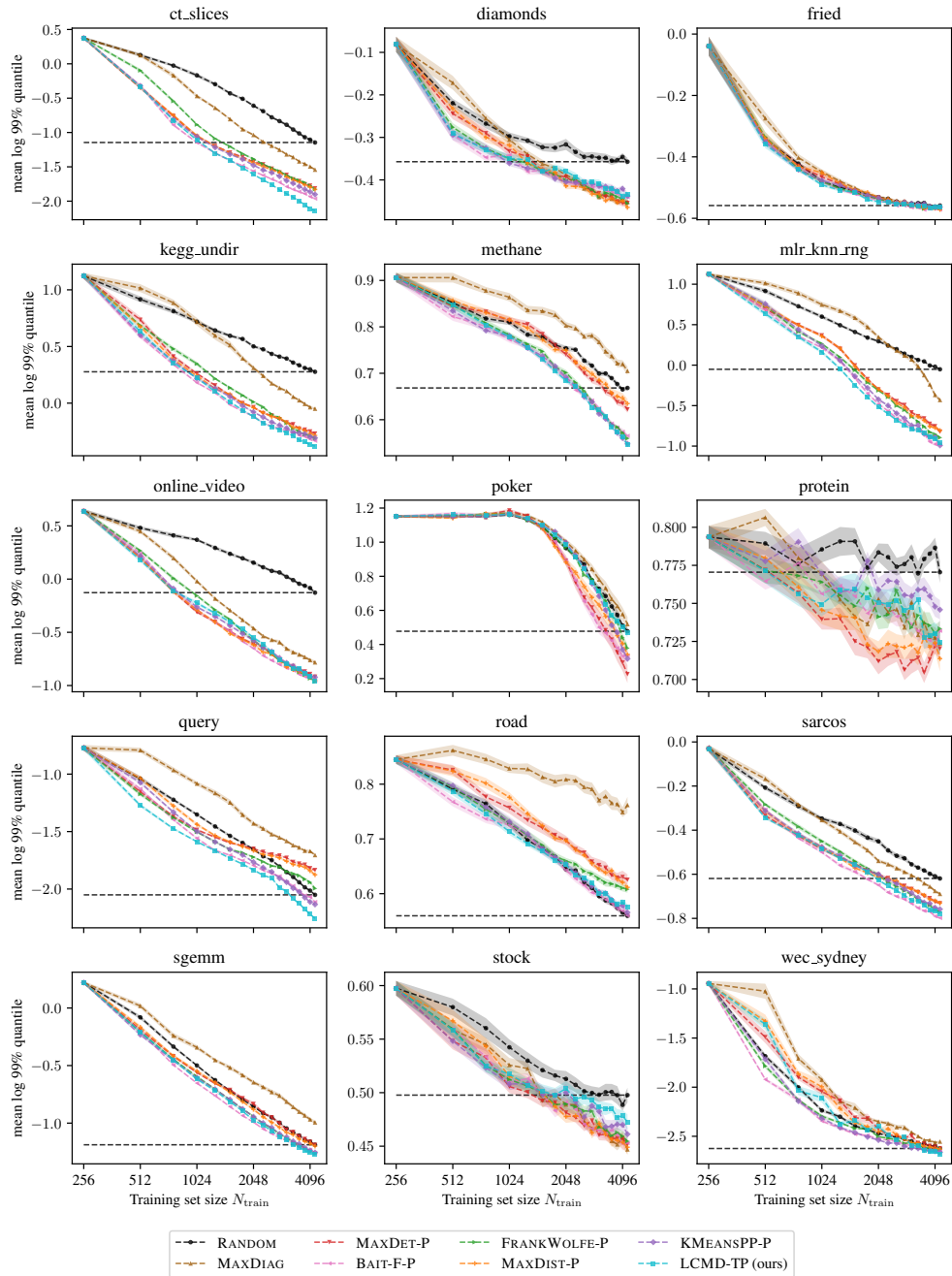


Figure E.2: This figure shows how fast the 99% quantile decreases during BMAL on the individual benchmark data sets for different selection methods and their corresponding kernels from Table 6. Specifically, the plots above show the logarithmic 99% quantile between each BMAL step for  $N_{\text{batch}} = 256$ , averaged over all repetitions. The black horizontal dashed line corresponds to the final performance of RANDOM at  $N_{\text{train}} = 4352$ . The shaded area corresponds to one estimated standard deviation of the mean estimator, cf. Section E.4.

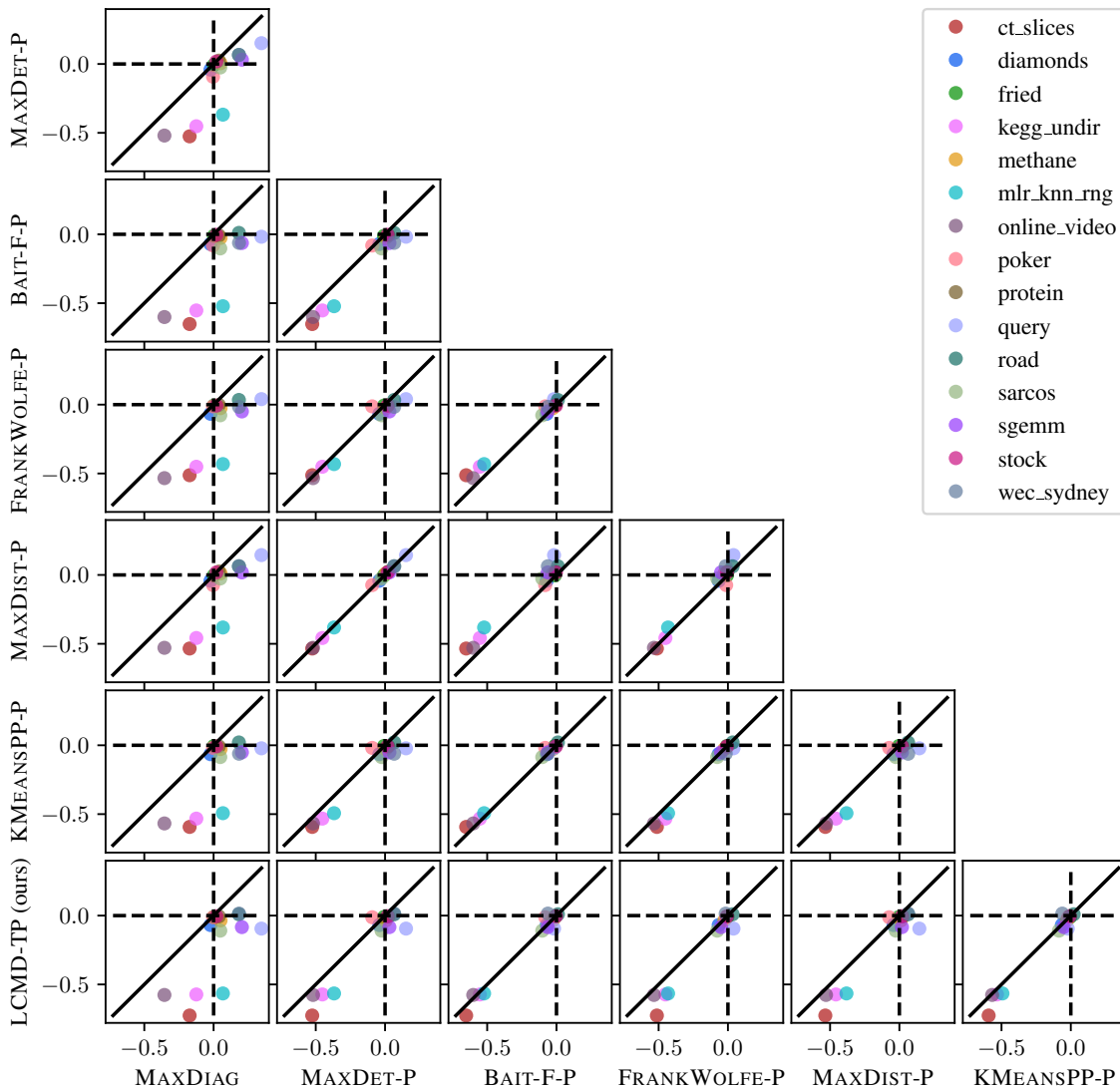


Figure E.3: Each subplot shows the errors of two selection methods and their corresponding selected kernels from Table 6. Specifically, the coordinates correspond to the mean log RMSE of the method on the data set minus the mean log RMSE of RANDOM selection on the same data set. Hence, the method on the  $x$  axis has a lower mean log RMSE than RANDOM on a data set if the corresponding point is left of the vertical dashed line, and it has a lower mean log RMSE than the method on the  $y$  axis if the corresponding point is left of the diagonal line. Similarly, the method on the  $y$  axis has a lower mean log RMSE than RANDOM on a data set if the corresponding point is below the horizontal dashed line, and it has a lower mean log RMSE than the method on the  $x$  axis if the corresponding point is below the diagonal line.

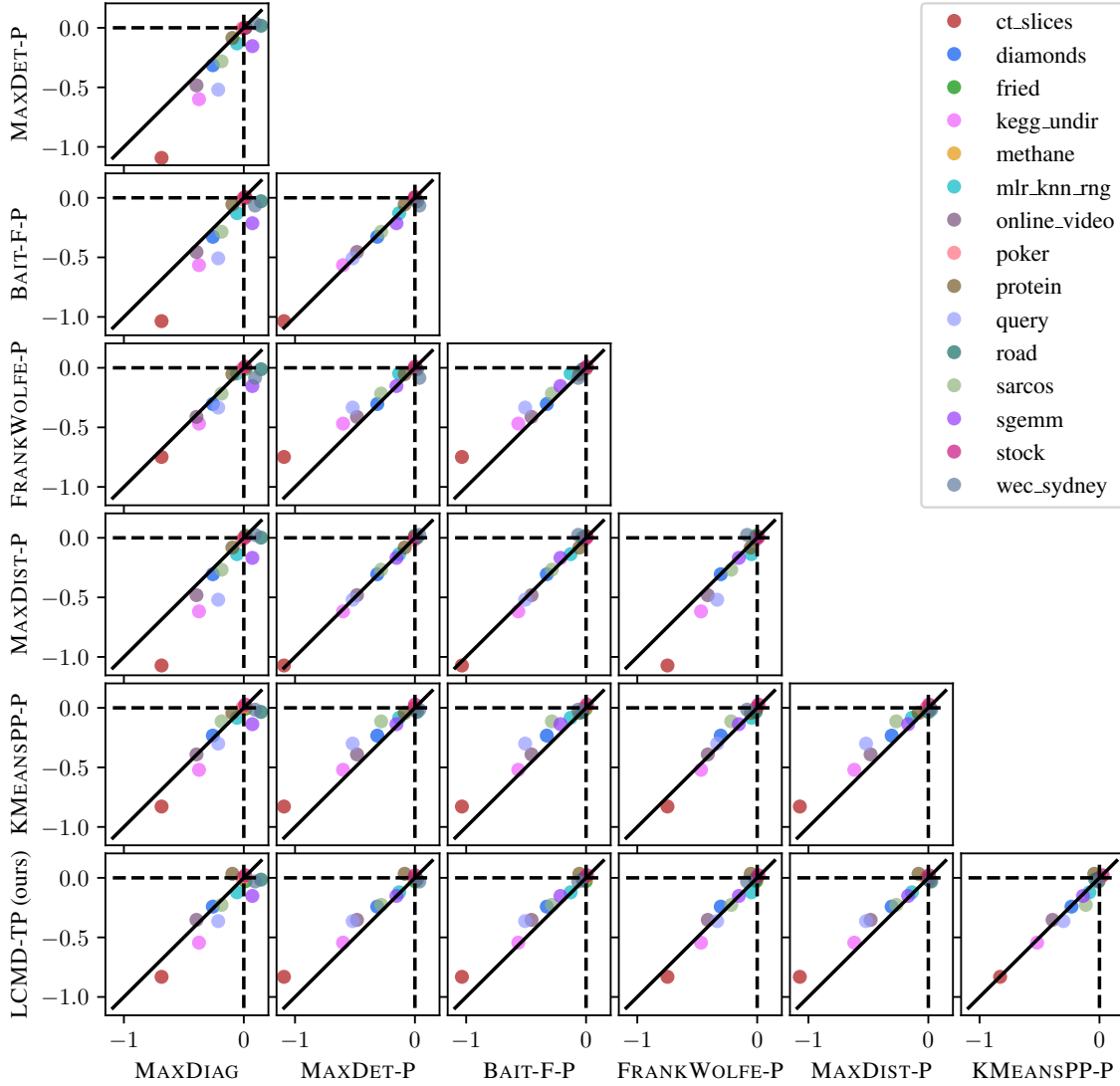


Figure E.4: Each subplot shows the errors of two selection methods and their corresponding selected kernels from Table 6. Specifically, the coordinates correspond to the mean log MAXE of the method on the data set minus the mean log MAXE of RANDOM selection on the same data set. Hence, the method on the  $x$  axis has a lower mean log MAXE than RANDOM on a data set if the corresponding point is left of the vertical dashed line, and it has a lower mean log MAXE than the method on the  $y$  axis if the corresponding point is left of the diagonal line.

## References

- Thomas D. Ahle, Michael Kapralov, Jakob BT Knudsen, Rasmus Pagh, Ameya Velingker, David P. Woodruff, and Amir Zandieh. Oblivious sketching of high-degree polynomial kernels. In *ACM-SIAM Symposium on Discrete Algorithms*, 2020.
- Rahaf Aljundi, Nikolay Chumerin, and Daniel Olmeda Reino. Identifying wrongly predicted samples: A method for active learning. In *Winter Conference on Applications of Computer Vision*, 2022.
- Christos Anagnostopoulos, Fotis Savva, and Peter Triantafillou. Scalable aggregation predictive analytics. *Applied Intelligence*, 48(9):2546–2567, 2018.
- Sanjeev Arora, Simon S. Du, Wei Hu, Zhiyuan Li, Russ R. Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. In *Neural Information Processing Systems*, 2019.
- Rosa I. Arriaga and Santosh Vempala. Algorithmic theories of learning. In *Foundations of Computer Science*, 1999.
- David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *ACM-SIAM Symposium on Discrete Algorithms*, 2007.
- Jordan Ash, Surbhi Goel, Akshay Krishnamurthy, and Sham Kakade. Gone fishing: Neural active learning with Fisher embeddings. In *Neural Information Processing Systems*, 2021.
- Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. In *International Conference on Learning Representations*, 2019.
- Alexander Atanasov, Blake Bordelon, and Cengiz Pehlevan. Neural networks as kernel learners: The silent alignment effect. In *International Conference on Learning Representations*, 2021.
- Rafael Ballester-Ripoll, Enrique G. Paredes, and Renato Pajarola. Sobol tensor trains for global sensitivity analysis. *Reliability Engineering & System Safety*, 183:311–322, 2019.
- Jörg Behler. Perspective: Machine learning potentials for atomistic simulations. *Journal of Chemical Physics*, 145(17):170901, 2016.
- William H. Beluch, Tim Genewein, Andreas Nürnberger, and Jan M. Köhler. The power of ensembles for active learning in image classification. In *Conference on Computer Vision and Pattern Recognition*, 2018.
- Marshall Bern and David Eppstein. Approximation algorithms for geometric problems. In *Approximation Algorithms for NP-hard Problems*, pages 296–345. PWS Publishing Company, 1996.
- Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.



- Zalán Borsos, Mojmir Mutny, and Andreas Krause. Coresets via bilevel optimization for continual learning and streaming. In *Neural Information Processing Systems*, 2020.
- Zalán Borsos, Marco Tagliasacchi, and Andreas Krause. Semi-supervised batch active learning via bilevel optimization. In *Conference on Acoustics, Speech and Signal Processing*, 2021.
- Erdem Bıyık, Kenneth Wang, Nima Anari, and Dorsa Sadigh. Batch active learning using determinantal point processes. *arXiv:1906.07975*, 2019.
- William F. Caselton and James V. Zidek. Optimal monitoring network designs. *Statistics & Probability Letters*, 2(4):223–227, 1984.
- M. Emre Celebi, Hassan A. Kingravi, and Patricio A. Vela. A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert Systems with Applications*, 40(1):200–210, 2013.
- Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, pages 273–304, 1995.
- Laming Chen, Guoxin Zhang, and Eric Zhou. Fast greedy map inference for determinantal point process to improve recommendation diversity. In *Neural Information Processing Systems*, 2018.
- Lenaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Neural Information Processing Systems*, 2019.
- Ali Civril and Malik Magdon-Ismail. Exponential inapproximability of selecting a maximum volume sub-matrix. *Algorithmica*, 65(1):159–176, 2013.
- David A. Cohn. Neural network exploration using optimal experiment design. *Neural Networks*, 9(6):1071–1083, 1996.
- Cody Coleman, Christopher Yeh, Stephen Mussmann, Baharan Mirzasoleiman, Peter Bailis, Percy Liang, Jure Leskovec, and Matei Zaharia. Selection via proxy: Efficient data selection for deep learning. In *International Conference on Learning Representations*, 2019.
- Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace Redux-Effortless Bayesian Deep Learning. In *Neural Information Processing Systems*, 2021.
- Stefano De Marchi, Robert Schaback, and Holger Wendland. Near-optimal data-independent point locations for radial basis function interpolation. *Advances in Computational Mathematics*, 23(3):317–330, 2005.
- Tewodros Deneke, Habtegebrel Haile, Sébastien Lafond, and Johan Lilius. Video transcoding time prediction for proactive load balancing. In *International Conference on Multimedia and Expo*, 2014.
- Dheeru Dua and Casey Graff. UCI Machine Learning Repository, 2017. URL <http://archive.ics.uci.edu/ml>.

- Stefan Elfving, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks*, 107:3–11, 2018.
- David Eppstein, Sarel Har-Peled, and Anastasios Sidiropoulos. Approximate greedy clustering and distance selection for graph metrics. *Journal of Computational Geometry*, 11(1):629–652, 2020.
- Runa Eschenhagen, Erik Daxberger, Philipp Hennig, and Agustinus Kristiadi. Mixtures of Laplace approximations for improved post-hoc uncertainty in deep learning. In *NeurIPS 2021 Workshop on Bayesian Deep Learning*, 2021.
- Sebastian Farquhar, Yarin Gal, and Tom Rainforth. On statistical bias in active learning: How and when to fix it. In *International Conference on Learning Representations*, 2021.
- Tomás Feder and Daniel Greene. Optimal algorithms for approximate clustering. In *ACM Symposium on Theory of Computing*, 1988.
- Valerii V. Fedorov. *Theory of Optimal Experiments*. Academic Press, New York, 1972.
- Stanislav Fort, Gintare Karolina Dziugaite, Mansheej Paul, Sepideh Kharaghani, Daniel M. Roy, and Surya Ganguli. Deep learning versus kernel learning: an empirical study of loss landscape geometry and the time evolution of the neural tangent kernel. In *Neural Information Processing Systems*, 2020.
- Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3(1-2):95–110, 1956.
- Jerome H. Friedman. Multivariate adaptive regression splines. *The Annals of Statistics*, pages 1–67, 1991.
- Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, 2016.
- Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International Conference on Machine Learning*, 2017.
- Yonatan Geifman and Ran El-Yaniv. Deep active learning over the long tail. *arXiv:1711.00941*, 2017.
- Amirata Ghorbani, James Zou, and Andre Esteva. Data shapley valuation for efficient batch active learning. In *Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2022.
- Teofilo F. Gonzalez. Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science*, 38:293–306, 1985.
- Yury Gorishniy, Ivan Rubachev, Valentin Khruikov, and Artem Babenko. Revisiting deep learning models for tabular data. In *Neural Information Processing Systems*, 2021.

- Franz Graf, Hans-Peter Kriegel, Matthias Schubert, Sebastian Pölsterl, and Alexander Cavallaro. 2D image registration in ct images using radial image descriptors. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2011.
- Insu Han, Haim Avron, Neta Shoham, Chaewon Kim, and Jinwoo Shin. Random features for the neural tangent kernel. *arXiv:2104.01351*, 2021.
- Lars Kai Hansen and Peter Salamon. Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(10):993–1001, 1990.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *IEEE Conference on Computer Vision*, 2015.
- Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv:1112.5745*, 2011.
- Alexander Immer, Maciej Korzepa, and Matthias Bauer. Improving predictions of Bayesian neural nets via local linearization. In *International Conference on Artificial Intelligence and Statistics*, 2021.
- Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural Tangent Kernel: Convergence and generalization in neural networks. In *Neural Information Processing Systems*, 2018.
- William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, 26, 1984.
- Arlind Kadra, Marius Lindauer, Frank Hutter, and Josif Grabocka. Well-tuned simple nets excel on tabular datasets. In *Neural Information Processing Systems*, 2021.
- Purushottam Kar and Harish Karnick. Random feature maps for dot product kernels. In *Artificial Intelligence and Statistics*, 2012.
- Ioannis Katsavounidis, C.-C. Jay Kuo, and Zhen Zhang. A new initialization technique for generalized Lloyd iteration. *IEEE Signal Processing Letters*, 1(10):144–146, 1994.
- Leonard Kaufman and Peter J. Rousseeuw. Finding groups in data: an introduction to cluster analysis. *Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics*, 1990.
- Manohar Kaul, Bin Yang, and Christian S. Jensen. Building accurate 3d spatial networks to enable next generation intelligent transportation systems. In *International Conference on Mobile Data Management*, 2013.
- Ronald W. Kennard and Larry A. Stone. Computer aided design of experiments. *Technometrics*, 11(1):137–148, 1969.
- Mohammad Emtiyaz E. Khan, Alexander Immer, Ehsan Abedi, and Maciej Korzepa. Approximate inference turns deep networks into gaussian processes. In *Neural Information Processing Systems*, 2019.

- Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.
- Andreas Kirsch, Joost Van Amersfoort, and Yarin Gal. BatchBALD: Efficient and diverse batch acquisition for deep Bayesian active learning. In *Neural Information Processing Systems*, 2019.
- Andreas Krause, Ajit Singh, and Carlos Guestrin. Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, 9(2), 2008.
- Agustinus Kristiadi, Matthias Hein, and Philipp Hennig. Being Bayesian, even just a bit, fixes overconfidence in relu networks. In *International Conference on Machine Learning*, 2020.
- Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.
- Anders Krogh and Jesper Vedelsby. Neural network ensembles, cross validation, and active learning. In *Neural Information Processing Systems*, 1994.
- Punit Kumar and Atul Gupta. Active learning query strategies for classification, regression, and clustering: a survey. *Journal of Computer Science and Technology*, 35(4):913–945, 2020.
- J. Nathan Kutz. Deep learning in fluid dynamics. *Journal of Fluid Mechanics*, 814:1–4, 2017.
- Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Neural Information Processing Systems*, volume 30, 2017.
- Pierre Simon Laplace. Mémoire sur la probabilité de causes par les évènements. *Mémoires de Mathématique et de Physique, Présentés à l’Académie Royale des Sciences, par divers Savants & lus dans ses Assemblées. Tome Sixième*, pages 621–656, 1774.
- Alexander Lavin, Hector Zenil, Brooks Paige, David Krakauer, Justin Gottschlich, Tim Mattson, Anima Anandkumar, Sanjay Choudry, Kamil Rocki, Atılım Güneş Baydin, and others. Simulation intelligence: Towards a new generation of scientific methods. *arXiv:2112.03235*, 2021.
- Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S. Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as Gaussian processes. In *International Conference on Learning Representations*, 2018.
- Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In *Neural Information Processing Systems*, 2019.

- Stuart Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.
- Philip M. Long. Properties of the after kernel. *arXiv:2105.10585*, 2021.
- Miguel Lázaro-Gredilla and Aníbal R. Figueiras-Vidal. Marginalized neural network mixtures for large-scale regression. *IEEE Transactions on Neural Networks*, 21(8):1345–1351, 2010.
- David JC MacKay. Bayesian interpolation. *Neural Computation*, 4(3):415–447, 1992a.
- David JC MacKay. Information-based objective functions for active data selection. *Neural Computation*, 4(4):590–604, 1992b.
- Vivek Madan, Mohit Singh, Uthaiapon Tantipongpipat, and Weijun Xie. Combinatorial algorithms for optimal design. In *Conference on Learning Theory*, 2019.
- Wesley J. Maddox, Pavel Izmailov, Timur Garipov, Dmitry P. Vetrov, and Andrew Gordon Wilson. A simple baseline for bayesian uncertainty in deep learning. In *Neural Information Processing Systems*, 2019.
- Alexander G. de G. Matthews, Jiri Hron, Mark Rowland, Richard E. Turner, and Zoubin Ghahramani. Gaussian process behaviour in wide deep neural networks. In *International Conference on Learning Representations*, 2018.
- Arash Mehrjou, Ashkan Soleymani, Andrew Jesson, Pascal Notin, Yarin Gal, Stefan Bauer, and Patrick Schwab. GeneDisco: A benchmark for experimental design in drug discovery. In *International Conference on Learning Representations*, 2021.
- Mohamad Amin Mohamadi, Wonho Bae, and Danica J. Sutherland. Making look-ahead active learning strategies feasible with Neural Tangent Kernels. In *Advances in Neural Information Processing Systems*, 2022.
- Douglas C. Montgomery. *Design and Analysis of Experiments*. John Wiley & Sons, 2017.
- Radford M. Neal. Priors for Infinite Networks. Technical Report CRG-TR-94-1, Dept. of Computer Science, University of Toronto, 1994.
- Mehdi Neshat, Bradley Alexander, Markus Wagner, and Yuanzhong Xia. A detailed comparison of meta-heuristic methods for optimising wave energy converter placements. In *Proceedings of the Genetic and Evolutionary Computation Conference*, 2018.
- Manuel Nonnenmacher, David Reeb, and Ingo Steinwart. Which minimizer does my neural network converge to? In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2021.
- Roman Novak, Jascha Sohl-Dickstein, and Samuel S. Schoenholz. Fast finite width neural tangent kernel. In *International Conference on Machine Learning*, 2022.
- Sebastian W. Ober and Carl Edward Rasmussen. Benchmarking the neural linear model for regression. In *Symposium on Advances in Approximate Bayesian Inference*, 2019.

- Rafail Ostrovsky, Yuval Rabani, Leonard J. Schulman, and Chaitanya Swamy. The effectiveness of Lloyd-type methods for the k-means problem. In *IEEE Symposium on Foundations of Computer Science*, 2006.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, and Luca Antiga. Pytorch: An imperative style, high-performance deep learning library. In *Neural Information Processing Systems*, 2019.
- Maryam Pazouki and Robert Schaback. Bases for kernel-based spaces. *Journal of Computational and Applied Mathematics*, 236(4):575–588, 2011.
- Robert Pinsler, Jonathan Gordon, Eric Nalisnick, and José Miguel Hernández-Lobato. Bayesian batch active learning as sparse subset approximation. In *Neural Information Processing Systems*, 2019.
- Remus Pop and Patric Fulop. Deep ensemble bayesian active learning: Addressing the mode collapse issue in monte carlo dropout via ensembles. *arXiv:1811.03897*, 2018.
- Maziar Raissi, Paris Perdikaris, and George E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019.
- Hiranmayi Ranganathan, Hemanth Venkateswara, Shayok Chakraborty, and Sethuraman Panchanathan. Deep active learning for image regression. In *Deep Learning Applications*, pages 113–135. Springer, Singapore, 2020.
- Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B. Gupta, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *ACM Computing Surveys*, 54(9):1–40, 2021.
- Daniel J. Rosenkrantz, Richard E. Stearns, and Philip M. Lewis. An analysis of several heuristics for the traveling salesman problem. *SIAM Journal on Computing*, 6(3):563–581, 1977.
- Gabriele Santin, Toni Karvonen, and Bernard Haasdonk. Sampling based approximation of linear functionals in reproducing kernel Hilbert spaces. *BIT Numerical Mathematics*, pages 1–32, 2021. Publisher: Springer.
- Fotis Savva, Christos Anagnostopoulos, and Peter Triantafillou. Explaining aggregates for exploratory analytics. In *International Conference on Big Data*, 2018.
- Nicol N. Schraudolph. Fast curvature matrix-vector products for second-order gradient descent. *Neural Computation*, 14(7):1723–1738, 2002.
- Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018.

- Sambu Seo, Marko Wallat, Thore Graepel, and Klaus Obermayer. Gaussian process regression: Active data selection and test point rejection. In *Mustererkennung 2000*, pages 27–34. Springer, 2000.
- Burr Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.
- H. Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Workshop on Computational Learning Theory*, 1992.
- Haozhe Shan and Blake Bordelon. A theory of neural tangent kernel alignment and its influence on training. *arXiv:2105.14301*, 2021.
- Claude Elwood Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.
- Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504, 2003.
- Apoorva Sharma, Navid Azizan, and Marco Pavone. Sketching curvature for efficient out-of-distribution detection for deep neural networks. In *Uncertainty in Artificial Intelligence*, 2021.
- Neta Shoham and Haim Avron. Experimental design for overparameterized learning with application to single shot deep active learning. *arXiv:2009.12820*, 2020.
- Kirstine Smith. On the standard deviations of adjusted and interpolated values of an observed polynomial function and its constants and the guidance they give towards a proper choice of the distribution of observations. *Biometrika*, 12(1/2):1–85, 1918.
- Jasper Snoek, Oren Rippel, Kevin Swersky, Ryan Kiros, Nadathur Satish, Narayanan Sundaram, Mostofa Patwary, Mr Prabhat, and Ryan Adams. Scalable bayesian optimization using deep neural networks. In *International Conference on Machine Learning*, 2015.
- Gowthami Somepalli, Micah Goldblum, Avi Schwarzschild, C. Bayan Bruss, and Tom Goldstein. SAINT: Improved neural networks for tabular data via row attention and contrastive pre-training. *arXiv:2106.01342*, 2021.
- Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.
- Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.
- Evgenii Tsymbalov, Maxim Panov, and Alexander Shapeev. Dropout-based active learning for regression. In *International Conference on Analysis of Images, Social Networks and Texts*, 2018.

- Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. OpenML: Networked science in machine learning. *SIGKDD Explorations*, 15(2):49–60, 2013.
- Sethu Vijayakumar and Stefan Schaal. Locally weighted projection regression: An  $o(n)$  algorithm for incremental real time learning in high dimensional space. In *International Conference on Machine Learning*, 2000.
- Abraham Wald. On the efficient design of statistical investigations. *The Annals of Mathematical Statistics*, 14(2):134–140, 1943.
- Haonan Wang, Wei Huang, Ziwei Wu, Hanghang Tong, Andrew J. Margenot, and Jingrui He. Deep active learning by leveraging training dynamics. In *Advances in Neural Information Processing Systems*, 2022.
- Zhilei Wang, Pranjali Awasthi, Christoph Dann, Ayush Sekhari, and Claudio Gentile. Neural active learning with performance guarantees. In *Neural Information Processing Systems*, 2021.
- Lilian Weng. Learning with not enough data part 2: Active learning. *lilianweng.github.io/lil-log*, 2022. URL <https://lilianweng.github.io/lil-log/2022/02/20/active-learning.html>.
- Tizian Wenzel, Gabriele Santin, and Bernard Haasdonk. A novel class of stabilized greedy kernel approximation algorithms: Convergence, stability and uniform point distribution. *Journal of Approximation Theory*, 262:105508, 2021.
- Andrew G. Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. In *Neural Information Processing Systems*, 2020.
- David P. Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends in Theoretical Computer Science*, 10(1–2):1–157, 2014.
- Dongrui Wu. Pool-based sequential active learning for regression. *IEEE Transactions on Neural Networks and Learning Systems*, 30(5):1348–1359, 2018.
- Henry P. Wynn. The sequential generation of D-optimum experimental designs. *The Annals of Mathematical Statistics*, 41(5):1655–1664, 1970.
- Hwanjo Yu and Sungchul Kim. Passive sampling for regression. In *International Conference on Data Mining*, 2010.
- Amir Zandieh, Insu Han, Haim Avron, Neta Shoham, Chaewon Kim, and Jinwoo Shin. Scaling neural tangent kernels via sketching and random features. In *Neural Information Processing Systems*, 2021.
- Viktor Zaverkin and Johannes Kästner. Exploration of transferable and uniformly accurate neural network interatomic potentials using optimal experimental design. *Machine Learning: Science and Technology*, 2(3):035009, 2021.
- Fedor Zhdanov. Diverse mini-batch active learning. *arXiv:1901.05954*, 2019.



Yilun Zhou, Adithya Renduchintala, Xian Li, Sida Wang, Yashar Mehdad, and Asish Ghoshal. Towards Understanding the Behaviors of Optimal Deep Active Learning Algorithms. In *International Conference on Artificial Intelligence and Statistics*, 2021.

Dominik Ślęzak, Marek Grzegorowski, Andrzej Janusz, Michał Kozielski, Sinh Hoa Nguyen, Marek Sikora, Sebastian Stawicki, and Łukasz Wróbel. A framework for learning and embedding multi-sensor forecasting models into a decision support system: A case study of methane concentration in coal mines. *Information Sciences*, 451:112–133, 2018.