The sample efficiency of Bayesian optimization (BO) is often boosted by Gaussian Process (GP) surrogate models. However, on mixed variable spaces, surrogate models other than GPs are prevalent, mainly due to the lack of kernels which can model complex dependencies across different types of variables. In this paper, we propose the frequency modulated (FM) kernel flexibly modeling dependencies among different types of variables, so that BO can enjoy the further improved sample efficiency. The FM kernel uses distances on continuous variables to modulate the graph Fourier spectrum derived from discrete variables. However, the frequency modulation does not always define a kernel with the similarity measure behavior which returns higher values for pairs of more similar points. Therefore, we specify and prove conditions for FM kernels to be positive definite and to exhibit the similarity measure behavior. In experiments, we demonstrate the improved sample efficiency of GP BO using FM kernels (BO-FM). On synthetic problems and hyperparameter optimization problems, BO-FM outperforms competitors consistently. Also, the importance of the frequency modulation principle is empirically demonstrated on the same problems. On joint optimization of neural architectures and SGD hyperparameters, BO-FM outperforms competitors including Regularized evolution (RE) and BOHB. Remarkably, BO-FM performs better even than RE and BOHB using three times as many evaluations.

1 INTRODUCTION

Bayesian optimization has found many applications ranging from daily routine level tasks of finding a tasty cookie recipe [Solnik et al., 2017] to sophisticated hyperparameter optimization tasks of machine learning algorithms (e.g. Alpha-Go [Chen et al., 2018]). Much of this success is attributed to the flexibility and the quality of uncertainty quantification of Gaussian Process (GP)-based surrogate models [Snoek et al., 2012, Swersky et al., 2013, Oh et al., 2018].

Despite the superiority of GP surrogate models, as compared to non-GP ones, their use on spaces with discrete structures (e.g., chemical spaces [Reymond and Awale, 2012], graphs and even mixtures of different types of spaces) is still application-specific [Kandasamy et al., 2018, Korovina et al., 2019]. The main reason is the difficulty of defining kernels flexible enough to model dependencies across different types of variables. On mixed variable spaces which consist of different types of variables including continuous, ordinal and nominal variables, current BO approaches resort to non-GP surrogate models, such as simple linear models or linear models with manually chosen basis functions [Daxberger et al., 2019]. However, such linear approaches are limited because they may lack the necessary model capacity.

There is much progress on BO using GP surrogate models (GP BO) for continuous, as well as for discrete variables. However, for mixed variables it is not straightforward how to define kernels which can model dependencies across different types of variables. To bridge the gap, we propose frequency modulation which uses distances on continuous variables to modulate the frequencies of the graph spectrum [Ortega et al., 2018] where the graph represents the discrete part of the search space [Oh et al., 2019].

A potential problem in the frequency modulation is that it does not always define a kernel with the similarity measure behavior [Vert et al., 2004]. That is, the frequency modulation does not necessarily define a kernel that returns higher values for pairs of more similar points. Formally, for a stationary kernel $k(x, y) = s(x - y)$, $s$ should be decreasing [Remes et al., 2017]. In order to guarantee the similarity measure behavior of kernels constructed by frequency modulation, we stipulate a condition, the frequency modula-
A crucial component in BO is thus the surrogate model. We coin frequency modulated (FM) kernels as the kernels constructed by frequency modulation and respecting the frequency modulation principle.

Different to methods that construct kernels on mixed variables by kernel addition and kernel multiplication, for example, FM kernels do not impose an independence assumption among different types of variables. In FM kernels, quantities in the two domains, that is the distances in a spatial domain and the frequencies in a Fourier domain, interact. Therefore, the restrictive independence assumption is circumvented, and thus flexible modeling of mixed variable functions is enabled.

In this paper, (i) we propose frequency modulation, a new way to construct kernels on mixed variables, (ii) we provide the condition to guarantee the similarity measure behavior of FM kernels together with a theoretical analysis, and (iii) we extend frequency modulation so that it can model complex dependencies between arbitrary types of variables. In experiments, we validate the benefit of the increased modeling capacity of FM kernels and the importance of the frequency modulation principle for improved sample efficiency on different mixed variable BO tasks. We also test BO with GP using FM kernels (BO-FM) on a challenging joint optimization of the neural architecture and the hyperparameters with two strong baselines, Regularized Evolution (RE) [Real et al., 2019] and BOHB [Falkner et al., 2018]. BO-FM outperforms both baselines which have proven their competence in neural architecture search [Dong et al., 2021]. Remarkably, BO-FM outperforms RE with three times evaluations.

2 PRELIMINARIES

2.1 BAYESIAN OPTIMIZATION WITH GAUSSIAN PROCESSES

Bayesian optimization (BO) aims at finding the global optimum of a black-box function $g$ over a search space $\mathcal{X}$. At each round BO performs an evaluation $y_i$ on a new point $x_i \in \mathcal{X}$, collecting the set of evaluations $\mathcal{D}_t = \{(x_i, y_i)\}_{i=1, \ldots, t}$ at the $t$-th round. Then, a surrogate model approximates the function $g$ given $\mathcal{D}_t$ using the predictive mean $\mu(x_i | \mathcal{D}_t)$ and the predictive variance $\sigma^2(x_i | \mathcal{D}_t)$. Now, an acquisition function $r(x_i | \mathcal{D}_t) = r(\mu(x_i | \mathcal{D}_t), \sigma^2(x_i | \mathcal{D}_t))$ quantifies how informative input $x \in \mathcal{X}$ is for the purpose of finding the global optimum. $g$ is then evaluated at $x_{t+1} = \arg\max_{x \in \mathcal{X}} r(x), y_{t+1} = g(x_{t+1})$. With the updated set of evaluations, $\mathcal{D}_{t+1} = \mathcal{D}_t \cup \{(x_{t+1}, y_{t+1})\}$, the process is repeated.

A crucial component in BO is thus the surrogate model. Specifically, the quality of the predictive distribution of the surrogate model is critical for balancing the exploration-exploitation trade-off [Shahriari et al. 2015]. Compared with other surrogate models (such as Random Forest [Hutter et al. 2011] and a tree-structured density estimator [Bergstra et al. 2011]), Gaussian Processes (GPs) tend to yield better results [Snoek et al. 2012; Oh et al. 2018].

For a given kernel $k$ and data $\mathcal{D} = \{(X, y)\}$ where $X = [x_1, \ldots, x_n]^T$ and $y = [y_1, \ldots, y_n]^T$, a GP has a predictive mean $\mu(x | \mathcal{D}) = k_X(k_X + \sigma^2)^{-1} y$ and predictive variance $\sigma^2(x | \mathcal{D}) = k_x - k_X(k_X + \sigma^2)^{-1} k_X$, where $k_x = k(x_i, x_i)$, $[k_X]_{ij} = k(x_i, x_j)$, $k_x = [k_X]^T$ and $[k_X]_{i,j} = k(x_i, x_j)$.

2.2 KERNELS ON DISCRETE VARIABLES

We first review some kernel terminology [Scholkopf and Smola 2001] that is needed in the rest of the paper.

Definition 2.1 (Gram Matrix). Given a function $k: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ and data $x_1, \ldots, x_n \in \mathcal{X}$, the $n \times n$ matrix $K$ with elements $[K]_{ij} = k(x_i, x_j)$ is called the Gram matrix of $k$ with respect to $x_1, \ldots, x_n$.

Definition 2.2 (Positive Definite Matrix). A real $n \times n$ matrix $K$ satisfying $\sum_i a_i [K]_{ij} a_j \geq 0$ for all $a_i \in \mathbb{R}$ is called positive definite (PD).

Definition 2.3 (Positive Definite Kernel). A function $k: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ which gives rise to a positive definite Gram matrix for all $n \in \mathbb{N}$ and all $x_1, \ldots, x_n \in \mathcal{X}$ is called a positive definite (PD) kernel, or simply a kernel.

A search space which consists of discrete variables, including both nominal and ordinal variables, can be represented as a graph [Kondor and Lafferty 2002; Oh et al. 2019]. In this graph each vertex represents one state of exponentially many joint states of the discrete variables. The edges represent relations between these states (e.g. if they are similar) [Oh et al. 2019]. With a graph representing a search space of discrete variables, kernels on a graph can be used for BO. In [Smola and Kondor 2003], for a positive decreasing function $f$ and a graph $G = (\mathcal{V}, \mathcal{E})$ whose graph Laplacian $L(G)$ has the eigendecomposition $UAU^T$, it is shown that a kernel can be defined as

$$k_{\text{disc}}(v, v' | \beta) = [U f(\Lambda | \beta) U^T]_{vv'}$$

(1)

where $\beta \geq 0$ is a kernel parameter and $f$ is a positive decreasing function. It is the reciprocal of a regularization operator [Smola and Kondor 2003] which penalizes high frequency components in the spectrum.

1Sometimes, different terms are used, semi-positive definite for $\sum_i a_i [K]_{ij} a_j \geq 0$ and positive definite for $\sum_i a_i [K]_{ij} a_j > 0$. Here, we stick to the definition in [Scholkopf and Smola 2001].

2In this paper, we use a (unnormalized) graph Laplacian $L(G) = D - A$ while, in [Smola and Kondor 2003], symmetric normalized graph Laplacian $L_{\text{sym}}(G) = D^{-1/2}(D - A)D^{-1/2}$. ($A$: adj. mat. / $D$: deg. mat.) Kernels are defined for both.
3  MIXED VARIABLE BAYESIAN OPTIMIZATION

With the goal of obtaining flexible kernels on mixed variables which can model complex dependencies across different types of variables, we propose the frequency modulated (FM) kernel. Our objective is to enhance the modeling capacity of GP surrogate models and, thereby improve the sample efficiency of mixed-variable BO. FM kernels use the continuous variables to modulate the frequencies of the kernel of discrete variables defined on the graph. As a consequence, FM kernels can model complex dependencies between continuous and discrete variables. Specifically, let us start with continuous variables of dimension $D_c$, and discrete variables represented by the graph $G = (\mathcal{V}, \mathcal{E})$ whose graph Laplacian $L(G)$ has eigenvalue decomposition $U A U^T$. To define a frequency modulated kernel we consider the function $k : (\mathbb{R}^{D_c} \times \mathcal{V}) \times (\mathbb{R}^{D_c} \times \mathcal{V}) \rightarrow \mathbb{R}$ of the following form

$$k((c,v),(c',v'))|\theta) = \sum_i |U|_{ij} f(\lambda_i, c - c' |\theta) |U|_{ij}$$

where $|c - c'|^2 = \sum c_d^2$ as arguments, and its output is combined with the basis $|U|_{ij}$. That is, the function $f$ processes the information in each eigencomponent separately while Eq. (2) then sums up the information processed by $f$. Note that unlike kernel addition and kernel product, the distance $|c - c'|^2$ influences each eigencomponent separately as illustrated in Figure 1. Unfortunately, Eq. (3) with an arbitrary function $f$ does not always define a positive definite kernel. Moreover, Eq. (2) with an arbitrary function $f$ may return higher kernel values for less similar points, which is not expected from a proper similarity measure \cite{Vert2004}. To this end, we first specify three properties of functions $f$ such that Eq. (2) guarantees to be a positive definite kernel and a proper similarity measure at the same time. Then, we motivate the necessity of each of the properties in the following subsections.

**Definition 3.1** (Frequency modulating function). A frequency modulating function is a function $f : \mathbb{R}^+ \times \mathbb{R} \rightarrow \mathbb{R}$ satisfying the three properties below.

**FM-P1** For a fixed $t \in \mathbb{R}$, $f(s,t)$ is a positive and decreasing function with respect to $s$ on $[0, +\infty)$.

**FM-P2** For a fixed $s \in \mathbb{R}^+$, $f(s, |c - c'| |\theta)$ is a positive definite kernel on $(c, c') \in \mathbb{R}^{D_c} \times \mathbb{R}^{D_c}$.

**FM-P3** For $t_1 < t_2$, $h_{t_1,t_2}(s) = f(s,t_1) - f(s,t_2)$ is positive, strictly decreasing and convex w.r.t $s \in \mathbb{R}^+$.\footnote{e.g. $k_{	ext{add}}((c,v),(c',v')) = e^{-|c-c'|^2} + k_{\text{disc}}(v,v')$ and $k_{	ext{prod}}((c,v),(c',v')) = e^{-|c-c'|^2} \cdot k_{\text{disc}}(v,v')$}

**3.1 FREQUENCY REGULARIZATION OF FM KERNELS**

In \cite{Smola2003}, it is shown that Eq. (1) defines a kernel that regularizes the eigenfunctions with high frequencies when $f$ is positive and decreasing. It is also shown that the reciprocal of $f$ in Eq. (1) is a corresponding regularization operator. For example, the diffusion kernel defined with $f(\lambda) = \exp(-\beta \lambda)$ corresponds to the regularization operator $r(\lambda) = \exp(\beta \lambda)$. The regularized Laplacian kernel defined with $f(\lambda) = 1/(1 + \beta \lambda)$ corresponds to the regularization operator $r(\lambda) = 1 + \beta \lambda$. Both regularization operators put more penalty on higher frequencies $\lambda$.

Therefore, the property FM-P1 forces FM kernels to have the same regularization effect of promoting a smoother function by penalizing the eigenfunctions with high frequencies.

**3.2 POSITIVE DEFINITENESS OF FM KERNELS**

Determining whether Eq.2 defines a positive definite kernel is not trivial. The reason is that the gram matrix $[k((c_i,v),(c_j,v_j))]_{i,j}$ is not determined only by the entries $v_i$ and $v_j$, but these entries are additionally affected by different distance terms $|c_i - c_j| |\theta$.

**Theorem 3.1.** If $f(\lambda, |c - c'| |\theta)$ defines a positive definite kernel with respect to $c$ and $c'$, then the FM kernel with such $f$ is positive definite jointly on $c$ and $v$. That is, the positive definiteness of $f(\lambda, |c - c'| |\theta)$ on $\mathbb{R}^{D_c}$ implies the positive definiteness of the FM kernel on $\mathbb{R}^{D_c} \times \mathcal{V}$.

**Proof.** See Supp. Sec. 1, Thm. 1.1

Note that Theorem 3.1 shows that the property FM-P2 guarantees that FMs kernels are positive definite jointly on $c$ and $v$. 
In the current form of Theorem 3.1, the frequency modulating functions depend on the distance $\|c - c'\|_\theta$. However, the proof does not change for the more general form of $f(\lambda, c, c' | \alpha, \beta)$, where $f$ does not depend on $\|c - c'\|_\theta$. Hence, Theorem 3.1 can be extended to the more general case that $f(\lambda, c, c' | \alpha, \beta)$ is positive definite on $(c, c') \in \mathbb{R}^{D_v} \times \mathbb{R}^{D_v}$.

## 3.3 Frequency Modulation Principle

A kernel, as a similarity measure, is expected to return higher values for pairs of more similar points and vice versa [Vert et al., 2004]. We call such behavior the similarity measure behavior.

In Eq. (2), the distance $\|c - c'\|_\theta$ represents a quantity in the "spatial" domain interacting with quantities $\lambda$,s in the "frequency" domain. Due to the interplay between the two different domains, the kernels of the form Eq. (2) do not exhibit the similarity measure behavior for an arbitrary function $f$. Next, we derive a sufficient condition on $f$ for the similarity measure behavior to hold for FM kernels.

Formally, the similarity measure behavior is stated as

$$\|c - c'\|_\theta \leq \|\hat{c} - \hat{c}'\|_\theta \Rightarrow k((c, v), (c', v')) \geq k((\hat{c}, v), (\hat{c}', v'))$$

or equivalently,

$$\|c - c'\|_\theta \leq \|\hat{c} - \hat{c}'\|_\theta \Rightarrow \sum_{i=1}^{\mid \mathcal{V} \mid} |U|_{i} h_{i, \alpha}(\lambda_{i} | \beta) |U|_{i, \beta} \geq 0 \tag{4}$$

where $h_{i, \alpha}(\lambda_{i} | \beta) = f(\lambda_{i}, 1_{i} | \beta) - f(\lambda_{i}, 1_{i} | \beta)$, $\alpha = \|c - c'\|_\theta$ and $\beta = \|\hat{c} - \hat{c}'\|_\theta$.

**Theorem 3.2.** For a connected and weighted undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with non-negative weights on edges, define a similarity (or kernel) $a(v, v') = [U \Lambda U^T]_{v, v'}$, where $U$ and $\Lambda$ are eigenvectors and eigenvalues of the graph Laplacian $L(\mathcal{G}) = U \Lambda U^T$. If $h$ is any non-negative and strictly decreasing convex function on $[0, \infty)$, then $a(v, v') \geq 0$ for all $v, v' \in \mathcal{V}$.

Therefore, these conditions on $h(\Lambda)$ result in a similarity measure $a$ with only positive entries, which in turn proves property Eq. (4). Here, we provide a proof of the theorem for a simpler case with an unweighted complete graph, where Eq. (4) holds without the convexity condition on $h$.

**Proof.** For an unweighted complete graph with $n$ vertices, we have eigenvalues $\lambda_1 = 0, \lambda_2 = \cdots = \lambda_n = n$ and eigenvectors such that $|U|_1 = 1/\sqrt{n}$ and $\sum_{i=1}^{n} |U|_{i} |U|_{i, \beta} = \delta_{\beta}$. For $v \neq v'$, the conclusion in Eq. (4), $\sum_{i=1}^{n} h(\lambda_i) |U|_{i, v} |U|_{i, v'}$ becomes $h(0)/n + h(n) \sum_{i=2}^{n} |U|_{i, v} |U|_{i, v'} = (h(0) - h(n))/n$ in which non-negativity follows with decreasing $h$.

For the complete proof, see Thm. 2.1 in Supp. Sec. 2.

**Corollary 3.2.1.** The random walk kernel derived from the symmetric normalized Laplacian [Smola and Kondor, 2003], the diffusion kernels [Kondor and Lafferty, 2002; Oh et al., 2019] and the regularized Laplacian kernel [Smola and Kondor, 2003] derived from symmetric normalized or unnormalized Laplacian, are all non-negative valued.

**Proof.** See Cor. 2.1 in Supp. Sec. 2.

## 3.4 FM Kernels In Practice

**Scalability** Since the (graph Fourier) frequencies and basis functions are computed by the eigendecomposition of cubic computational complexity, a plain application of frequency modulation makes the computation of $1$ kernels prohibitive for a large number of discrete variables. Given $P$ discrete variables where each variable can be individually represented by a graph $\mathcal{G}_P$, the discrete part of the search space can be represented as a product space, $\mathcal{V} = \mathcal{V}_1 \times \cdots \times \mathcal{V}_P$.

In this case, we define FM kernels on $\mathbb{R}^{D_v} \times \mathcal{V} = \mathbb{R}^{D_v} \times (\mathcal{V}_1 \times \cdots \times \mathcal{V}_P)$ as

$$k((c, v), (c', v')) = \prod_{p=1}^{P} [U(V)_{p, c} | \alpha, \beta, \theta] = \prod_{p=1}^{P} \sum_{i=1}^{\mid \mathcal{V}_p \mid} [U(V)_{p, c} | \alpha, \beta, \theta]$$

where $v = (v_1, \cdots, v_P), v' = (v'_1, \cdots, v'_P), \alpha = (\alpha_1, \cdots, \alpha_P)$, $\beta = (\beta_1, \cdots, \beta_P)$ and the graph Laplacian is given as $L(\mathcal{G}_P)$ with the eigendecomposition $U_{p} \Lambda_{p} U_{p}^T$. Eq.5 should not be confused with the kernel product of kernels on each $\mathcal{V}_p$. Note that the distance $\|c - c'\|_\theta$ is shared, which introduces the coupling among discrete variables and thus allows more modeling freedom than a product kernel. In addition to the coupling, the kernel parameter $\alpha$s lets us individually determine the strength of the frequency modulation.

**Examples** Defining a FM kernel amounts to constructing a frequency modulating function. We introduce examples of flexible families of frequency modulating functions.
Proposition 1. For $S \in (0, \infty)$, a finite measure $\mu$ on $[0, S]$, $\mu$-measurable $\tau : [0, S] \rightarrow [0, 2]$ and $\mu$-measurable $\rho : [0, S] \rightarrow \mathbb{N}$, the function of the form below is a frequency modulating function.

$$f(\lambda, c||e-c'||\theta|\beta) = \int_{S}^{1} \frac{1}{(1 + \beta\lambda + \alpha||e-c'||^{2})^\rho(s)^\rho(s)} \mu(ds)$$

(6)


Assuming $S = 1$ and $\tau(s) = 2$, Prop. 1 gives $(1 + \beta\lambda + \alpha||e-c'||^{2})^{-1}$ with $\rho(s) = 1$ and $\mu(ds) = ds$, and $\sum_{n=1}^{\infty} a_n (1 + \beta\lambda + \alpha||e-c'||^{2})^{-n}$ with $\rho(s) = \lfloor Ns \rfloor$ and $\mu(\lfloor Ns \rfloor) = a_n \geq 0$ and $\mu(\lfloor n/N \rfloor_{n=1,\ldots,N}) = 0$.

3.5 EXTENSION OF THE FREQUENCY MODULATION

Frequency modulation is not restricted to distances on Euclidean spaces but it is applicable to any arbitrary space with a kernel defined on it. As a concrete example of frequency modulation by kernels, we show a non-stationary extension a kernel defined on it. As a concrete example of frequency modulation by kernels, we show a non-stationary extension of the frequency modulating function.

$$\sum_{n=1}^{\infty} a_n (1 + \beta\lambda + \alpha||e-c'||^{2})^{-n}$$

where $a_n$ is a similarity measure and thus the property $FM-P3$ is satisfied. If the premise $t_1 < t_2$ of the property $FM-P3$ is replaced by $t_1 > t_2$, then $FM-P3$ is also satisfied. In contrast to the frequency modulation principle with distances in Eq. (3), the frequency modulation principle with a kernel is formalized as

$$k_{NN}(e, c'||\Sigma) \leq k_{NN}(e, c||\Sigma)$$

(8)

Note that $k_{NN}(e, c||\Sigma)$ is a similarity measure and thus the inequality is not reversed unlike Eq. (3).

All above arguments on the extension of the frequency modulation using a nonstationary kernel hold also when the $k_{NN}$ is replaced by an arbitrary positive definite kernel. The only required condition is that a kernel has to be upper bounded, i.e., $k_{NN}(e, c') \leq C$, needed for $FM-P1$ and $FM-P2$.

4 RELATED WORK

On continuous variables, many sophisticated kernels have been proposed [Wilson and Nickisch, 2015, Samo and Roberts, 2015, Remes et al., 2017, Oh et al., 2018]. In contrast, kernels on discrete variables have been studied less [Haussler, 1999, Kondor and Lafferty, 2002, Smola and Kondor, 2003]. To our best knowledge, most of existing kernels on mixed variables are constructed by a kernel product [Swersky et al., 2013, Li et al., 2016] with some exceptions [Krause and Ong, 2011, Swersky et al., 2013, Fiducioso et al., 2019], which rely on kernel addition.

In mixed variable BO, non-GP surrogate models are more prevalent, including SMAC [Hutter et al., 2011] using random forest and TPE [Bergstra et al., 2011] using a tree structured density estimator. Recently, by extending the approach of using Bayesian linear regression for discrete variables [Baptista and Poloczek, 2018, Daxberger et al., 2019] proposes Bayesian linear regression with manually chosen basis functions on mixed variables, providing a regret analysis using Thompson sampling as an acquisition function. Another family of approaches utilizes a bandit framework to handle the acquisition function optimization on mixed variables with theoretical analysis [Gopakumar et al., 2018, Nguyen et al., 2019, Ru et al., 2020]. Nguyen et al. [2019] proposes Bayesian linear regression with manually chosen basis functions on mixed variables, providing a regret analysis using GP-UCB. Among these approaches, Ru et al. [2020] also utilize information across different categorical values, which—in combination with the bandit framework—makes itself the most competitive method in the family.

Our focus is to extend the modeling prowess and flexibility of pure GPs for surrogate models on problems with mixed variables. We propose frequency modulated kernels, which are kernels that are specifically designed to model the complex interactions between continuous and discrete variables.

In architecture search, approaches using weight sharing such as DARTS [Liu et al., 2018] and ENAS [Pham et al., 2018] are gaining popularity. In spite of their efficiency, methods training neural networks from scratch for given architectures are under-explored with a few exceptions such as BOHB [Falkner et al., 2018] and autoHAS [Dong et al., 2019]. Our approach proposes a competitive option to this challenging optimization of mixed variable functions with expensive evaluation cost.

5 EXPERIMENTS

To demonstrate the improved sample efficiency of GP BO using FM kernels (BO-FM) we study various mixed variable black-box function optimization tasks, including 3 synthetic
problems from [Ru et al., 2020], 2 hyperparameter optimization problems (SVM [Smola and Kondor, 2003] and XGBoost [Chen and Guestrin, 2016]) and the joint optimization of neural architecture and SGD hyperparameters. For synthetic problems and hyperparameter experiments are provided in Supp. Sec. 4.

In each round, after updating with an evaluation, we fit a GP surrogate model using marginal likelihood maximization with 10 random initialization until convergence [Rasmussen, 2003]. We use the expected improvement (EI) acquisition function [Donald, 1998] and optimize it by repeated alternation of L-BFGS-B [Zhu et al., 1997] and hill climbing [Skiena, 1998] until convergence. More details on the experiments are provided in Supp. Sec. 4.

Baselines For synthetic problems and hyperparameter optimization problems below, baselines we consider are SMAC [Hutter et al., 2011], TPE [Bergstra et al., 2011], and CoCaBO [Ru et al., 2020] which consistently outperforms One-hot BO [authors, 2016] and EXP3BO [Gopakumar et al., 2018]. For CoCaBO, we consider 3 variants using different mixture weights. 

5.1 SYNTHETIC PROBLEMS

We test on 3 synthetic problems proposed in [Ru et al., 2020][10]. Each of the synthetic problems has the search space as in Tab. 1. Details of synthetic problems can be found in [Ru et al., 2020].

Table 1: Synthetic Problem Search Spaces

<table>
<thead>
<tr>
<th>Conti. Space</th>
<th>Num. of Cats.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Func2C</td>
<td>$[-1,1]^2$</td>
</tr>
<tr>
<td>Func3C</td>
<td>$[-1,1]^2$</td>
</tr>
<tr>
<td>Ackley5C</td>
<td>$[-1,1]$</td>
</tr>
</tbody>
</table>

5.2 HYPERPARAMETER OPTIMIZATION PROBLEMS

Now we consider a practical application of Bayesian optimization over mixed variables. We take two machine learn-
We optimize hyperparameters of NuSVR in scikit-learn [Pedregosa et al. 2011]. We consider 3 categorical hyperparameters and 3 continuous hyperparameters (Tab. 2) and for continuous hyperparameters we search over log_{10} transformed space of the range.

### NuSVR hyperparameters

<table>
<thead>
<tr>
<th>NuSVR param</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>kernel</td>
<td>{linear, poly, RBF, sigmoid}</td>
</tr>
<tr>
<td>gamma</td>
<td>{scale, auto}</td>
</tr>
<tr>
<td>shrinking</td>
<td>{on, off}</td>
</tr>
<tr>
<td>C</td>
<td>[10^{-3}, 10]</td>
</tr>
<tr>
<td>tol</td>
<td>[10^{-6}, 1]</td>
</tr>
<tr>
<td>nu</td>
<td>[10^{-6}, 1]</td>
</tr>
</tbody>
</table>

Table 2: NuSVR hyperparameters

For each of 5 split of Boston housing dataset with train:test(7:3) ratio, NuSVR is fitted on the train set and RMSE on the test set is computed. The average of 5 test RMSE is the objective.

### XGBoost

We consider 1 ordinal, 3 categorical and 4 continuous hyperparameters (Tab. 3).

<table>
<thead>
<tr>
<th>XGBoost param</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>max_depth</td>
<td>{1, · · · , 10}</td>
</tr>
<tr>
<td>booster</td>
<td>{gbtree, dart}</td>
</tr>
<tr>
<td>grow_policy</td>
<td>{depthwise, lossguide}</td>
</tr>
<tr>
<td>objective</td>
<td>{multi:softmax, multi:softprob}</td>
</tr>
<tr>
<td>eta</td>
<td>[10^{-5}, 1]</td>
</tr>
<tr>
<td>gamma</td>
<td>[10^{-4}, 10]</td>
</tr>
<tr>
<td>subsample</td>
<td>[10^{-3}, 1]</td>
</tr>
<tr>
<td>lambda</td>
<td>[0, 5]</td>
</tr>
</tbody>
</table>

Table 3: XGBoost hyperparameters

For 3 continuous hyperparameters, eta, gamma and subsample, we search over the log_{10} transformed space of the range.

Comparison to different kernel combinations In Supp. Sec. 5 we also report the comparison with different kernel combinations on all 3 synthetic problems and 2 hyperparameter parameter optimization problems. We make two observations. First, MODDIF, which does not respect the similarity measure behavior, sometimes severely degrades BO performance. Second, MODLAP obtains equally good final results and consistently finds the better solutions faster than the kernel product. This can be clearly shown by comparing the area above the mean curve of BO runs using different kernels. The area above the mean curve of BO using MODLAP is larger than the are above the mean curve of BO using the kernel product. Moreover, the gap between the area from MODLAP and the area from kernel product increases in problems with larger search spaces. Even on the smallest search space, Func2C, MODLAP lags behind the kernel product up to around 90th evaluation and outperforms after it. The benefit of MODLAP modeling complex dependency among mixed variables is more prominent in higher dimension problems.

Ablation study on regression tasks In addition to the results on BO experiments, we compare FM kernels with kernel addition and kernel product on three regression tasks from UCI datasets (Supp. Sec. 6). In terms of negative log-likelihood (NLL), which takes into account uncertainty, ModLap performs the best in two out of three tasks. Even on the task which is conjectured to have a structure suitable to kernel product, ModLap shows competitive performance. Moreover, on regression tasks, the importance of the frequency modulation principle is further reinforced. For full NLL and RMSE comparison and detailed discussion, see Supp. Sec. 6.


![Figure 3: SVM(left), XGBoost(right) (Mean±Std.Err. of 5 runs)](image)
5.3 **JOINT OPTIMIZATION OF NEURAL ARCHITECTURE AND SGD HYPERPARAMETERS**

Next, we experiment with BO on mixed variables by optimizing continuous and discrete hyperparameters of neural networks. The space of discrete hyperparameters \( \mathcal{A} \) is modified from the NASNet search space [Zoph and Le, 2016], which consists of 8,153,726,976 choices. The space of continuous hyperparameters \( \mathcal{H} \) comprises 6 continuous hyperparameters of the SGD with a learning rate scheduler: learning rate, momentum, weight decay, learning rate reduction factor, 1st reduction point ratio and 2nd reduction point ratio. A good neural architecture should both achieve low errors and be computationally modest. Thus, we optimize the objective \( f(a, h) = err_{val}(a, h) + 0.02 \times \text{FLOP}(a)/\max_{a' \in \mathcal{A}} \text{FLOP}(a') \). To increase the separability among smaller values, we use log \( f(a, h) \) transformed values whenever model fitting is performed on evaluation data. The reported results are still the original non-transformed \( f(a, h) \).

We compare with two strong baselines. One is BOHB [Falkner et al., 2018] which is an evaluation-cost-aware algorithm augmenting unstructured bandit approach [Li et al., 2017] with model-based guidance. Another is RE [Real et al., 2019] based on a genetic algorithm with a novel population selection strategy. In Dong et al. [2021], on discrete-only spaces, these two outperform competitors including weight sharing approaches such as DARTS [Liu et al., 2018], SETN [Dong and Yang, 2019], ENAS [Pham et al., 2018] and etc. In the experiment, for BOHB, we use the public implementation [13] and for RE, we use our own implementation.

For a given set of hyperparameters, with ModLap or RE, the neural network is trained on FashionMNIST for 25 epochs while BOHB adaptively chooses the number of epochs. For further details on the setup and the baselines we refer the reader to Supp. Sec. 4 and 5.

We present the results in Fig. 4. Since BOHB adaptively chooses the budget (the number of epochs), BOHB is plotted according to the budget consumption. For example, the y-axis value of BOHB on 100-th evaluation is the result of BOHB having consumed 2,500 epochs (25 epochs \( \times \) 100).

We observe that ModLap finds the best architecture in terms of accuracy and computational cost. What is more, we observe that ModLap reaches the better solutions faster in terms of numbers of evaluations. Even though the time to evaluate a new hyperparameter is dominant, the time to suggest a new hyperparameter in ModLap is not negligible in this case. Therefore, we also provide the comparison with respect the wall-clock time. It is estimated that RE and BOHB evaluate 230 hyperparameters while ModLap evaluate 200 hyperparameters (Supp. Sec. 7). For the same estimated wall-clock time, ModLap(200) outperforms competitors (RE(230), BOHB(230)).

In order to see how beneficial the sample efficiency of BO-FM is in comparison to the baselines, we perform a stress test in which more evaluations are allowed for RE and BOHB. We leave RE and BOHB for 600 evaluations. Notably, ModLap with 200 evaluations outperforms both competitors with 600 evaluations (Fig. 4 and Supp. Sec. 5). We conclude that ModLap exhibits higher sample efficiency than the baselines.

### Table 1: Comparison with baselines

<table>
<thead>
<tr>
<th>Method</th>
<th>#Eval.</th>
<th>Mean±Std.Err.</th>
</tr>
</thead>
<tbody>
<tr>
<td>BOHB</td>
<td>200</td>
<td>( 7.158 \times 10^{-2} \pm 1.030 \times 10^{-3} )</td>
</tr>
<tr>
<td>BOHB</td>
<td>230</td>
<td>( 7.151 \times 10^{-2} \pm 9.836 \times 10^{-4} )</td>
</tr>
<tr>
<td>BOHB</td>
<td>600</td>
<td>( 6.941 \times 10^{-2} \pm 4.432 \times 10^{-4} )</td>
</tr>
<tr>
<td>RE</td>
<td>200</td>
<td>( 7.067 \times 10^{-2} \pm 1.141 \times 10^{-3} )</td>
</tr>
<tr>
<td>RE</td>
<td>230</td>
<td>( 7.061 \times 10^{-2} \pm 1.129 \times 10^{-3} )</td>
</tr>
<tr>
<td>RE</td>
<td>400</td>
<td>( 6.929 \times 10^{-2} \pm 6.480 \times 10^{-4} )</td>
</tr>
<tr>
<td>RE</td>
<td>600</td>
<td>( 6.879 \times 10^{-2} \pm 1.003 \times 10^{-3} )</td>
</tr>
<tr>
<td>ModLap</td>
<td>200</td>
<td>( 6.850 \times 10^{-2} \pm 3.791 \times 10^{-4} )</td>
</tr>
</tbody>
</table>

For the figure with all numbers above, see Supp. Sec. 5.

---

[13]: https://github.com/automl/HpBandSter

6 **CONCLUSION**

We propose FM kernels to improve the sample efficiency of mixed variable Bayesian optimization.

On the theoretical side, we provide and prove conditions for FM kernels to be positive definite and to satisfy the similarity measure behavior. Both conditions are not trivial due to the interactions between quantities on two disparate...
domains, the spatial domain and the frequency domain.

On the empirical side, we validate the effect of the conditions for FM kernels on multiple synthetic problems and realistic hyperparameter optimization problems. Further, we successfully demonstrate the benefits of FM kernels compared to non-GP based Bayesian Optimization on a challenging joint optimization of neural architectures and SGD hyperparameters. BO-FM outperforms its competitors, including Regularized evolution, which requires three times as many evaluations.

We conclude that an effective modeling of dependencies between different types of variables improves the sample efficiency of BO. We believe the generality of the approach can have a wider impact on modeling dependencies between discrete variables and variables of arbitrary other types, including continuous variables.

References


