arXiv:2303.01774v1 [cs.LG] 3 Mar 2023

通过基于字典的嵌入进行高维组合空间的贝叶斯优化

Aryan Deshwal
华盛顿州立大学

Sebastian Ament
Meta

Maximilian Balandat
Meta

Eytan Bakshy
Meta

Janardhan Rao Doppa
华盛顿州立大学

David Eriksson
Meta

摘要

我们研究如何在高维组合空间中优化计算成本高昂的黑盒函数，这类问题广泛存在于科学、工程和机器学习应用中。我们采用贝叶斯优化（BO），并提出一种新颖的代理建模方法，以高效处理大量二元和类别参数。其核心思路是从输入空间中选取若干离散结构（即字典），并利用它们为高维组合结构定义序数嵌入。这样，我们便可以使用现有的适用于连续空间的高斯过程模型。我们提出一种基于二元小波的系统化方法，用于构建二元空间的字典，并提出一种可推广至类别空间的随机构建方法。我们提供了理论依据，以支持基于字典的嵌入的有效性。在多个不同的真实世界基准测试中，我们的实验表明，与最先进的贝叶斯优化方法相比，所提出的代理建模方法效果显著。

1 引言

许多现实世界的应用都需要在有限的训练数据下，针对高维离散输入空间和混合输入空间（同时包含离散参数和连续参数）构建概率模型。这些模型需要对未知输入做出准确预测，并量化其不确定性，以及一些示例，包括环境模型的校准、用于自动机器学习（AutoML）的特征选择（其中某个特征的纳入或排除可表示为一个二元参数），以及微生物组分析（其中某个微生物物种的纳入或排除是一个二元参数，而环境变量对应于连续参数）。

高斯过程（GP） [Rasmussen， 2004]非常适用于这种情形。高斯过程也常用作代理模型，以样本高效地优化连续空间和离散／混合空间中的昂贵黑盒函数 [Frazier， 2018]。例如，在微生物组设计优化中，我们需要进行昂贵的湿实验室实验，以评估每种混合配置；这类配置由候选微生物子集和环境条件构成 [Clark 等人， 2021]。其他应用示例包括机器学习模型的特征选择（Guyon 和 Elisseeff [2003], ）、通过调整编译器标志来优化效率（Hellsten 等人 [2022], ），以及调整数据库配置（Zhang 等人 [2021]）。将高斯过程用于组合空间的关键挑战在于，如何定义合适的核函数来捕捉输入对之间的相似性。

本文提出了一种基于字典的新型汉明嵌入（HED）。这种嵌入使我们能够利用具有自动相关性判定（ARD）的常用高斯过程核函数，对高维组合输入进行建模。通过使用乘积核，我们的方法可以自然地扩展到同时包含连续变量和离散变量的混合输入。我们建模方法的核心思想是从输入空间中选取固定数量的候选结构，称为字典，并利用输入与字典元素之间的汉明距离，为输入空间定义一种嵌入 。这种方法的有效性在很大程度上取决于字典的选择。我们的理论分析表明，在 HED 上训练的高斯过程老虎机的遗憾界 [Srinivas 等人，2010] 是嵌入搜索空间基数的函数，而该基数又取决于一种正交性度量

发表于 26th2023 年人工智能与统计学国际会议（AISTATS），西班牙瓦伦西亚。PMLR：第 206 卷。版权所有 © 2023 作者。

字典。我们发现，构造字典时若能在初期限制搜索空间基数的缩减，就能获得很强的建模灵活性，从而实现由数据驱动的搜索空间压缩，并在实证中加快收敛。受这些理论见解启发，我们提出两种构造字典的方法：1）子采样二进小波 [Swanson 和 Tewfik, 1996], ，该方法在二的幂次维度上优化正交性度量；2）一种可推广至类别型输入的随机方法，并允许我们设计任意大小的字典。

为评估基于字典的嵌入的有效性，我们在贝叶斯优化（BO）框架下考虑了若干代价高昂的黑盒优化问题。将 BO 应用于组合空间面临着独特的挑战 [Doppa, 2021] 因为常用的代理模型在这种情况下往往表现不佳，而且我们无法依赖基于梯度的方法来优化效用函数。已应用于组合空间的代理模型包括采用扩散核的高斯过程 [Oh 等人， 2019], 采用各向同性核的高斯过程 [Wan 等人， 2021], 线性模型 [Baptista 和 Poloczek, 2018], 以及随机森林 [Hutter 等人, 2011年，Deshwal 等人， 2020]。在贝叶斯优化中使用基于字典的嵌入时，我们将其称为采用字典的贝叶斯优化（BODi）。我们在贝叶斯优化基准测试上进行了全面评估，结果表明 BODi 优于最先进的方法，并提供了实证证据，说明 BODi 的出色性能源于基于字典的代理模型。

本文的核心贡献是开发并评估我们基于字典的建模方法。我们的具体贡献包括：

1. 一种基于字典的嵌入，可显著提升高维组合空间和混合输入空间中高斯过程模型的质量。2. 两种构建字典的方法：1）通过二元小波；2）一种可推广至任意维度和类别变量的随机方法。3. 对我们方法的理论分析表明，在某些条件下，它能够压缩输入空间的基数；若采用自动相关性确定，则压缩效果取决于数据。4. 压缩后的基数可改善二元输入高斯过程老虎机的遗憾界。5. 在种类丰富的组合空间和混合空间贝叶斯优化基准上进行的全面实验评估，展现了 BODi 的有效性。源代码见 https://github.com/aryandeshwal/BODi。

2 背景

组合空间和混合空间。 设 $\mathcal{Z}$ 为一个组合空间，其中每个元素$\mathbf{z} \in \mathcal{Z}$ 都是一种离散结构。我们假设$\mathbf{z} \in \mathcal{Z}$ 可以用d 个离散变量 $v_1, v_2, \cdots, v_d$ 表示，其中每个变量 v_i的取值来自一个有限候选集 $C(v_i)$。每个变量 v_i 有 $\tau_i \geq 2$ 种可能取值，空间的基数为 $|\mathcal{Z}| = \prod_{i=1}^{d} \tau_i$。特别地，对于二元空间，$C(v_i) = \{0, 1\}$ 对所有 v_i 和 $|\mathcal{Z}| = 2^d$均成立。若 $\mathcal{X}$ 是一个由连续参数构成的空间，我们称 $\mathcal{X} \times \mathcal{Z}$ 为混合空间。

问题定义。 给定一个高维组合空间 $\mathcal{Z}$，即离散变量 d 的数量很大。我们假设要优化一个黑盒目标函数 f：$\mathcal{Z} \mapsto \mathbb{R}$，并且可以对每种结构$\mathbf{z} \in \mathcal{Z}$进行评估。例如，在自动机器学习任务的特征选择中，$\mathbf{z}$ 是一种二元结构，对应于一个特征子集，而 $f(\mathbf{z})$表示使用所选特征训练的机器学习模型的性能。我们的目标是在函数评估次数较少的情况下，找到一个近似优化 f 的结构$\mathbf{z} \in \mathcal{Z}$。

贝叶斯优化。 贝叶斯优化方法根据以往函数评估的训练数据构建概率代理模型 $\mathcal{M}$（通常为高斯过程），并以高效利用采样的方式，智能地选择待评估输入的序列。输入的选择通过最大化采集函数 α 来完成，该函数基于代理模型提供的后验分布进行计算。将贝叶斯优化用于高维组合空间的一项关键挑战是构建准确的代理模型，这正是我们工作的核心关注点。

高斯过程。高斯过程是一类非参数概率模型，因其灵活性和出色的不确定性量化能力而广受欢迎。高斯过程由均值函数和协方差函数或核函数指定 k：$\mathbb{R} \times \mathbb{R} \to \mathbb{R}$ [Rasmussen, 2004]。一种常见选择是径向基函数核或平方指数核，其定义为

$$k(\mathbf{x}, \mathbf{y}) = s^2 \exp\Big\{-\tfrac{1}{2} \sum_i (x_i - y_i)^2 / \ell_i^2\Big\}$$

其中， ℓ_i 对于 $i = 1, \cdots, D$ 是用于自动相关性确定（ARD）的长度尺度， s^2 是信号方差。

3 相关工作

离散与混合空间。近年来，针对离散与混合空间的贝叶斯优化因其在科学、工程、自动机器学习及其他领域的广泛适用性而备受关注。针对低维场景，人们提出了多种代理模型，

但正如我们的实验所示，这些模型通常不适用于高维空间。

BOCS [Baptista 和 Poloczek 2018] 针对二元空间，并采用二阶贝叶斯线性回归替代模型，该模型在输入维度增加时扩展性较差，可能无法支持底层黑盒函数需要更复杂模型的应用。已有工作还考虑了高斯过程模型的不同实例化形式。COMBO [Oh 等人，2019年、Deshwal 等人，2021a]在输入空间的组合图表示上，采用带有离散扩散核的高斯过程。最近，Kim 等人 [2022] 提出了一种针对组合空间的连续嵌入方法。该方法与我们的方法不同之处在于：他们采用均匀随机单射，并且需要在嵌入空间中优化采集函数后重构离散输入。此外，还有研究利用深度生成模型创建潜在空间，并应用连续贝叶斯优化方法（通常称为“潜空间贝叶斯优化”）：G´omez-Bombarelli 等人、[2018], Tripp 等人、[2020],Eissman 等人、[2018], Kajino、[2019], Notin 等人、[2021],Deshwal 和 Doppa、[2021], Maus 等人 [2022]。相比之下，我们的基于字典的嵌入计算效率高，并且利用了组合空间中固有的结构。探索如何协同结合潜在空间与基于字典的嵌入的优势，是一个富有价值的研究方向。

此外，也有先前研究探讨如何为同时包含离散变量和连续变量的混合空间构建核函数 [Ru 等人，2020年，Oh 等人，2021年，Deshwal 等人，2021b]。Garrido-Merch´án 和 Hern´ández-Lobato [2020] 会先对输入变量进行舍入，然后将其传入采用典范核的高斯过程。树结构帕岑估计器（TPEs） [Bergstra 等人， 2011] 适用于混合空间，并会在输入空间中进行密度估计，这在高维场景下可能颇具挑战。SMAC [Hutter 等人， 2011]采用随机森林代理模型。

高维连续空间。 关于高维连续空间上的贝叶斯优化，已有大量研究成果，可分为以下几类：(1) 低维结构，包括随机嵌入 [Wang 等人，2016 年，Letham 等人，2020年，Papenmeier 等人， 2022]，基于哈希的方法 [Nayebi 等人， 2019]，稀疏性诱导先验 [Eriksson 和 Jankowiak， 2021]，或学习得到的嵌入 [Garnett 等人， 2013]。(2) 加性结构 [Kandasamy 等人，2015年，Gardner 等人， 2017]，该方法假设高维黑盒函数可以分解为多个低维函数之和。(3) 避免选择不确定性很高的边界点的方法。Eriksson 等人 [2019] 使用以最优解为中心的局部信赖域，并根据进展情况调整这些信赖域的大小。Kirschner 等人 [2019] 沿一维直线优化采集函数，而 Oh 等人 [2018] 使用圆柱核来聚焦于区域内部。

然而，这些方法专用于连续空间，而对于许多现实应用中出现的高维组合搜索空间和混合搜索空间所面临的挑战，相关研究寥寥无几。一个例外是近期提出的 CASMOPOLITAN 方法 [Wan 等人， 2021]，该方法采用连续空间中的自适应信赖域 [Eriksson 等人，2019]并以汉明距离取代标准欧几里得距离，从而适用于离散（子）空间。我们提出的 BODi 算法及其配套的基于字典的核函数，在高维场景下优于 CASMOPOLITAN。

4 字典嵌入

在本节中，我们介绍通过**汉明嵌入**字典（HED）实现嵌入的思想。这是一种针对二元和类别型输入的新型嵌入方法，可将输入嵌入到有序特征空间中。具体而言，我们在基于字典$\mathbf{A}$ 的嵌入 $\boldsymbol{\phi}_{\mathbf{A}}(\mathbf{z})$ 上使用高斯过程，其中字典$\mathbf{A}$ 包含来自输入空间 $\mathcal{Z}$的m 个离散 d维元素。通过计算 h 与字典$\mathbf{a}_i \in \mathbf{A}$中每个元素之间的汉明距离，得到大小为 m 的嵌入 $\boldsymbol{\phi}_{\mathbf{A}}(\mathbf{z})$。也就是说，

$$[\boldsymbol{\phi}_{\mathbf{A}}(\mathbf{z})]_i = h(\mathbf{a}_i, \mathbf{z}).$$

HED 有几个优点。首先，它使我们能够将构建高维离散空间模型这一颇具挑战性的任务，转化为在成熟的连续空间设定下应用高斯过程。由此，我们可以推断与嵌入表示相关的长度尺度；相比之下，在原始类别空间中，一个长度尺度只能建模单个类别变化的影响。此外，嵌入使长度尺度的推断更为高效，从而能够通过自动相关性确定（ARD）有效剔除冗余维度；我们证明，这会降低输入空间的基数。我们从理论上证明，这提高了高斯过程老虎机（置信上界，UCB）的样本效率；此类方法常用于贝叶斯优化（BO）。该方法还在高维组合空间上的贝叶斯优化中取得了最先进的结果。此外，尽管核心核函数适用于二元空间，但只需使用乘积核，就能轻松将其扩展到同时包含连续参数和离散参数的混合空间。

字典构建过程。 HED 的有效性取决于字典构造。一个朴素的方法是简单地从二元空间中均匀分布地随机方法选取元素。然而，这种朴素方法在本文所考虑的测试

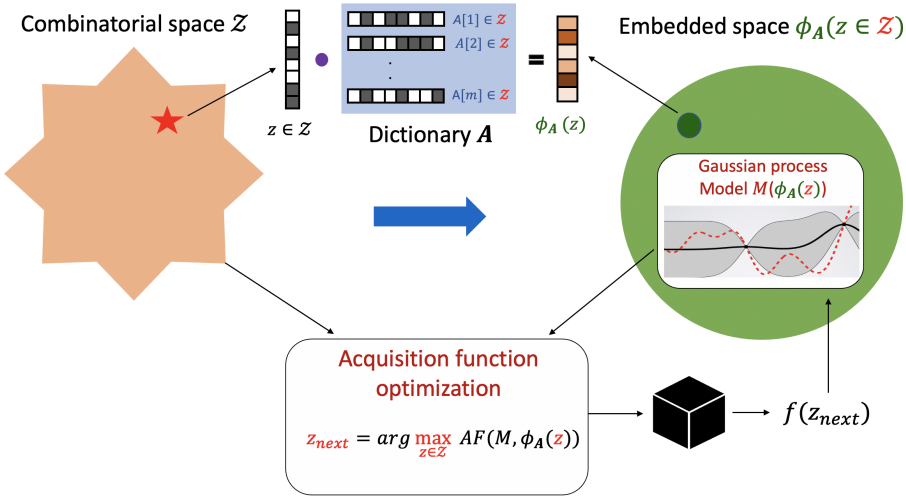
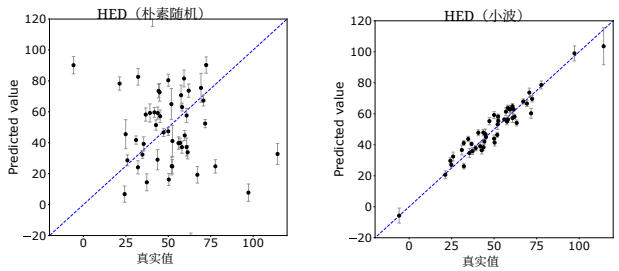


图 1：BODi 算法在二元空间中的高层概览。字典 $\mathbf{A}$ 包含来自组合空间 $\mathcal{Z}$ 的 m 个离散结构。每个高维二元结构$\mathbf{z} \in \mathcal{Z}$（以黑白方块表示）都会嵌入到低维嵌入空间 $\phi_{\mathbf{A}}(\mathbf{z}) \in \mathbb{R}^m$（以彩色方块表示）中。我们在嵌入空间上学习高斯过程代理模型，并在原始组合空间 $\mathcal{Z}$ 中执行采集函数优化，以便在每次贝叶斯优化迭代中选择下一个结构$\mathbf{z}_{next}$进行函数评估。



(a) 朴素随机字典 (b) 二元小波字典

图 2：在包含 60 个二元变量的最大可满足性问题上，比较朴素随机字典（左）与二元小波字典（右）的均值预测及其对应的 95% 预测区间（详见第 7 节）；使用 50 个训练点，并在 50 个测试点上进行预测。

问题上表现出较差的预测性能或贝叶斯优化性能。例如，图 2a 展示了在一个包含60个二元变量的 MaxSAT 测试问题上，使用具有均匀分布随机二元字典的字典核的高斯过程所表现出的较差预测性能。

另一种思路是采用确定性字典构造方法，例如多分辨率小波 [Mallat， 1989],这是一种有效且广为人知的工具，可通过对数据应用一组正交变换来研究不同尺度下的实值信号。在二元空间的背景下，二进制小波变换 [Swanson 和 Tewfik， 1996] 与著名的正交 Hadamard 矩阵高度相关，并被应用于信号处理、光谱学和密码学 [Hedayat 和 Wallis，1978，Horadam，2012]。与朴素的随机字典相比，子-采样的二进制小波字典在同一最大可满足性问题上实现了出色的预测性能，如图2b所示。

虽然二元小波为二元搜索空间中的预测与优化问题提供了强大的字典设计（相关优化结果见图 7），但对于非二的幂次维度，其构造并不简单；而且，任意维度下二元小波是否存在仍是一个开放问题 [Hadamard，1893年，Baumert等人，1962年，Djoković等人， 2014]。因此，在非二的幂次维度的实验中，我们对二的幂次维度的二元小波列进行子采样，详情见附录 C。

为克服二元小波一般构造上的困难，并将我们的方法推广到类别空间，我们提出了一种随机化过程，可生成稀疏度范围较大的字典行。我们将这种随机化过程称为“多样随机”。

算法 1 给出了用于在二元输入空间$\{0,1\}^d$上构造多样随机字典的伪代码。该构造过程的核心原则是通过生成由伯努利分布不同偏置参数（θ）决定的二进制向量，使字典行呈现多样性；这与朴素随机方法不同，在后者中，θ 始终为 $1/2$。因此，随着 d 增大，朴素随机字典的行往往趋近于含有 $d/2$ 个非零元素，而由于 θ 各不相同，多样随机字典则呈现出较大的稀疏度范围。该算法可以轻松推广到包含不同基数类别变量的输入；详情见附录 G。总而言之，

多样随机字典可以针对任意维度构造，也可以自然地扩展到类别型输入；正如我们稍后将展示的，它在广泛的基准问题上表现出优异的优化性能。

算法 1 二元输入空间的字典设计
$\{0,1\}^d$ 具有稀疏程度各异的行

需要：字典大小 m
1：字典 $\mathbf{A}$ ←为空
2: **对于** i=1, 2, . . . , m **执行**
3: 一个$_i$ ←空
4: 采样伯努利参数 $\theta \sim$ 均匀分布(0,1)
5: **对于** j=1, 2, . . . , d **执行**
6: 采样二进制数 $a \sim$ 伯努利(θ)
7: $\mathbf{a}_i \leftarrow \mathbf{a}_i \cup a$
8: **结束循环**
9: 将$\mathbf{a}_i$ 添加到字典中：$\mathbf{A} \leftarrow \mathbf{A} \cup \mathbf{a}_i$
10: **结束循环**
11: **返回**大小为 $m \times d$ 的字典 $\mathbf{A}$

混合输入空间的表示。我们主要关注纯组合输入空间 $\mathcal{Z}$，但可以很自然地将我们的方法扩展到由离散参数和连续参数组成的混合搜索空间。在这种情况下，我们旨在对输入空间 $\mathcal{X} \times \mathcal{Z}$ 进行建模，其中 $\mathcal{X}$ 是连续参数的定义域。为了将我们的方法扩展到这种混合输入设置，我们使用乘积核函数：对离散参数采用 HED 嵌入，对连续参数则采用带有自动相关性确定（ARD）的标准核函数，例如 Mat´ér-$5/2$ 核函数。

5 BODi：基于字典嵌入的贝叶斯优化

我们提出的 BODi 方法是通用贝叶斯优化框架的一种直接实例化。我们采用带有自动相关性确定（ARD）的标准 Mat´ern-$5/2$ 核函数，在 HED 嵌入上构建高斯过程（GP）代理模型，并采用单目标问题中常用的期望改进（EI）采集函数。在我们的设定中，EI 将代理模型 $\mathcal{M}$ 和嵌入 $\phi_{\mathbf{A}}(\mathbf{z})$作为输入，用于评估结构$\mathbf{z} \in \mathcal{Z}$的效用。为了在离散空间$\mathcal{Z}$上优化采集函数，我们采用从随机生成的初始条件出发的局部搜索。

算法 2 展示了我们方法的伪代码。我们使用一组较小的随机初始训练集，其中包含 $\mathcal{Z}$ 中的元素及其函数评估，以构建初始代理模型 $\mathcal{M}(\phi_{\mathbf{A}}(\mathbf{z}))$。在每次贝叶斯优化迭代中，我们使用算法 1 中描述的随机化过程生成新的字典 $\mathbf{A}$ ，并使用相应的嵌入 $\phi_{\mathbf{A}}(\mathbf{z})$重新拟合高斯过程模型。对于每次贝叶斯优化迭代 j，我们通过优化采集函数来选择下一个结构 $\mathbf{z}_j$ 。我们将 $\mathbf{z}_j$ 及其对应的函数值 $f(\mathbf{z}_j)$ 添加到训练数据 D_j 中，并使用 D_j训练新的代理模型$\mathcal{M}(\phi_{\mathbf{A}}(\mathbf{z}))$。我们重复这些步骤，直到查询预算耗尽，然后返回最佳输入$\mathbf{z}_{\text{best}} \in \mathcal{Z}$。

算法 2BODi (m) 算法

需要：黑箱目标函数 f，离散空间 $\mathcal{Z}$，其维数为 d，字典大小为 m
1: $D_0 \leftarrow$ 小规模随机初始训练数据
2: **对于** j=1、2、.....**执行**
3: 构建大小为 m 的字典$\mathbf{A}$
4: 为每个输入结构计算低维嵌入 $\phi_{\mathbf{A}}(\mathbf{z})$
使用字典 $\mathbf{A}$计算$\mathbf{z} \in D_j$
5: 在嵌入空间 $\phi_{\mathbf{A}}(\mathbf{z})$上拟合高斯过程 $\mathcal{M}$
6: 在离散空间中最大化采集函数
：$\mathbf{z}_j = \arg\max_{\mathbf{z}\in\mathcal{Z}} \alpha(\mathcal{M}(\phi_{\mathbf{A}}(\mathbf{z})))$
7: 评估选定的结构 $\mathbf{z}_j$ 以获得 $f(\mathbf{z}_j)$
8: 汇总训练数据：$D_j \leftarrow D_{j-1} \cup \{\mathbf{z}_j,\ f(\mathbf{z}_j)\}$
9: **结束循环**
10: **返回** $\mathbf{z}_{best} = \arg\min\{f(\mathbf{z}_1)\ f(\mathbf{z}_2)\ \cdots\}$

为了在混合搜索空间上优化采集函数，我们在连续子空间和离散子空间上交替搜索，这是贝叶斯优化在混合空间中的常见方法 [Oh 等人，2021年，Deshwal 等人，2021b，Wan 等人， 2021]。我们对离散参数使用局部搜索，对连续参数使用基于梯度的优化。尽管在离散空间上优化采集函数颇具挑战，但实践证明，采用重启的局部搜索是有效的 [Oh 等人， 2019]。

6 BODi 的理论分析

下面我们推导出汉明嵌入与仿射变换之间一个出人意料的关系，从而解释为什么典范线性嵌入（例如高斯嵌入）表现不佳。我们还为使用字典核的贝叶斯优化提供了一个遗憾界，该结果关键依赖于嵌入搜索空间基数的减少，而非维度的降低。我们的结果针对二元搜索空间给出，但利用二进制编码（例如独热编码），或更高效地使用 $\lceil \log_2(c) \rceil$ 位表示 c 个类别，可以直接推广到类别变量。

我们的第一个命题表明，$\{0,1\}^d$ 中向量的汉明嵌入等价于二元向量的 $\{\pm 1\}$ 编码的仿射变换。

命题 1（仿射表示）。令 $\mathbf{A} \in \{0,1\}^{m\times d}$，$\mathbf{z} \in \{0,1\}^d$。则

$$2\phi_{\mathbf{A}}(\mathbf{z}) = d\mathbf{1}_m - \bar{\mathbf{A}}\bar{\mathbf{z}}, \tag{1}$$

where $\bar{a}_{ij} = 2a_{ij} - 1$ *and* $\bar{z}_i = 2z_i - 1 \in \{-1, 1\}$.

证明。 参见附录A。 □

将式（1）代入嵌入距离公式可得

$$2\|\boldsymbol{\phi}_{\mathbf{A}}(\mathbf{z}) - \boldsymbol{\phi}_{\mathbf{A}}(\mathbf{z}')\|_2 = \|\bar{\mathbf{A}}\bar{\mathbf{r}}\|_2,$$

其中 $\bar{}\mathbf{r} = \bar{}\mathbf{z} - \bar{}\mathbf{z}'$。也就是说，距离计算仅依赖于对二进制输入向量的$\{\pm 1\}$-编码之差向量 $\bar{}\mathbf{r}$ 的线性 投影。此外，附录 C 中描述的小波字典所对应的嵌入，除一个常数偏移外，等价于子采样哈达玛变换，这是一种定义在布尔域上的傅里叶变换。

命题 1 证明，基于字典的核函数等价于在经过线性投影的输入数据上评估的典范核函数（例如 Mat´ern）。鉴于此前关于子空间上的贝叶斯优化 [Wang 等人，2016 年, Letham 等人，2020] 以及线性投影性质 [Larsen 和 Nelson， 2017], 的大量研究，人们或许会认为，像高斯随机矩阵这样的典范线性嵌入设计在我们的设定中会表现良好。然而，事实并非如此，我们将在实证评估中对此进行展示。

要理解其中的原因，首先要注意，BODi 实际上是在变换后的搜索空间中进行优化

$$\mathcal{S}_{\mathbf{A}} = \left\{\boldsymbol{\phi}_{\mathbf{A}}(\mathbf{z}) \mid \mathbf{z} \in \{0,1\}^d\right\}.$$

虽然线性嵌入通常会降低搜索空间的维度，但不一定会减少基数 $|\mathcal{S}_{\mathbf{A}}|$。基数是有限搜索空间中贝叶斯优化遗憾界的一个关键量。事实上，尽管高斯随机投影具有许多理想性质，包括近似保持距离和降低维度，但我们的下一个结果表明，即使是一维高斯随机投影，几乎必然也会保留原始搜索空间的全部基数。

命题 2。定义 $\mathcal{S}_{\mathbf{a}} = \{\mathbf{a}^\top \mathbf{z} \mid \mathbf{z} \in \{\pm 1\}^d\}$，并令$\mathbf{a} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$。则 $|\mathcal{S}_{\mathbf{a}}| = 2^d$ 几乎必然成立。

证明。参见附录 B。 □

相比之下，我们的下一个结果给出了 $\mathcal{S}_{\mathbf{A}}$ 基数的一个上界，该上界取决于字典行的变异性度量 $\mu_{\mathbf{A}}$，且仅随 d呈多项式增长。

命题 3 (嵌入基数**)**。 设$\mathbf{A} \in \{0,1\}^{m\times d}$。那么，嵌入搜索空间 $\mathcal{S}_{\mathbf{A}}$ 的基数上界为

$$|\mathcal{S}_{\mathbf{A}}| \le [(\mu_{\mathbf{A}}+1)(d+1-\mu_{\mathbf{A}})]^{\lfloor m/2 \rfloor}(d+1)^{m \bmod 2}$$

其中 $\mu_{\mathbf{A}} = \max_{i,\, j} \max(h(\mathbf{a}_i, \mathbf{a}_j)\ h(\neg\mathbf{a}_i,\ \mathbf{a}_j))$，且 h 是汉明距离。

证明。 参见附录 B。 □

命题 1 的仿射表示表明， $\mu_{\mathbf{A}}$ 与字典行的相干性具有很强的相似性:

$$2\mu_{\mathbf{A}} = d + \max_{i,j} \left|\bar{\mathbf{a}}_i^\top \bar{\mathbf{a}}_j\right|.$$

字典列的互相干性是压缩感知理论中的一个核心量 [Tropp, 2004]。此外， $\mu_{\mathbf{A}}$ 为字典设计提供了理论依据。事实上，附录 C 中的二元小波字典在二的幂次维度下达到了可能的最低互相干性 $d/2$ ，并在多种基准测试中表现出色（见图 7）。直观而言，我们希望将搜索空间的基数缩小到足以加速优化，但又不能缩小太多，以免它无法成为有用的归纳偏置。注意，$d/2 \le \mu_{\mathbf{A}} \le d$ ，且该上界在 $\mu_{\mathbf{A}} = d/2$ 时达到最大值。例如，字典中存在重复元素意味着$\mu_{\mathbf{A}} = d$，与附录 C 中的小波字典相比，在相同的 m 下会导致基数大幅下降。

现在，我们准备结合命题 3 的界与 Srinivas 等人 [2010] 的经典结果，为 BODi 提供改进的遗憾界。回顾一下，第 t次迭代时的遗憾定义为 $r_t = f(\mathbf{z}^*) - f(\mathbf{z}_t)$，其中，$\mathbf{z}^*$ 是最优点，而$\mathbf{z}_t$ 是在第 t^{th} 次迭代中选取的点。累积遗憾为 $R_T = \sum_{t=1}^{T} f(\mathbf{z}^*) - f(\mathbf{z}_t)$，是贝叶斯优化算法理论研究中的一个关键量。许多贝叶斯优化方法都是无遗憾的（即 $\lim_{T\to\infty} R_T/T = 0$），尽管 R_T趋近于零的速率存在显著差异。

Srinivas 等人 [2010]证明了一个关于 T的次线性遗憾界，该界适用于采用置信上界（UCB）采集函数 $\arg\max_{\mathbf{z}} \mu_{t-1}(\mathbf{z}) + \sqrt{\beta_t}\sigma_{t-1}(\mathbf{z})$的基于高斯过程的优化，其中 μ_t （分别为 σ_t^2）是高斯过程经过 t 次迭代后的预测均值（分别为方差）。该界主要取决于两个量：（1）经过 T次迭代后的信息增益 $\gamma_T = \log|\mathbf{I} + \sigma^{-2}\mathbf{K}_T|$，其中$\mathbf{K}_T$是根据前 T 次迭代中选取的输入 $\{\mathbf{z}_t\}_{t=1}^T$ 计算得到的核矩阵， σ 是观测噪声的标准差。（2）搜索空间 $|\mathcal{S}|$的基数，我们在命题 3 中对 BODi 的该基数进行了界定。值得注意的是， γ_T 取决于核函数；对于我们实验中的 Mat´ern-ν 核， $\gamma_T = \mathcal{O}(T^{d(d+1)/(2\nu + d(d+1))}\log T)$。下文中，我们将使用$\mathcal{O}^*$ 指代 $\mathcal{O}$，并省略对数因子。

定理 4. 设$\mathbf{A}$ 有 m 行， $\delta \in (0,\ 1)$，且 $\beta_t = 2\log(|\mathcal{S}_{\mathbf{A}}|t^2\pi^2/6\delta)$。那么，对于使用核函数为 $k_{BODi}(\mathbf{z},\ \mathbf{z}') = k_{base}(\boldsymbol{\phi}_{\mathbf{A}}(\mathbf{z})\ \boldsymbol{\phi}_{\mathbf{A}}(\mathbf{z}'))$的零均值高斯过程进行 f 次采样并运行置信上界算法，其累积遗憾以概率 $1-\delta$ 上界为 $\mathcal{O}^*(\sqrt{T\gamma_T m})$，其中 γ_T是 k_{base} 的最大信息增益。

定理 4 展现出依赖维数的遗憾标度降低至 $\mathcal{O}^*(\sqrt{m})$；相比之下，对于非嵌入式二进制输入，其遗憾标度为 $\bar{\mathcal{O}}^*(\sqrt{d})$，前提是 m 不太大。我们

强调，这归因于搜索空间基数的压缩，而非嵌入维度的降低。不过，还需注意，优化性能并非只取决于基数，因为通常存在两个相互矛盾的主要目标：(1) 找到一个表达能力足够强的模型；(2) 降低拟合和优化该模型的复杂度。单纯缩小搜索空间的基数虽然会使模型更容易拟合，但也可能降低模型准确刻画底层黑盒目标函数的可能性。

使用较大的字典，可以让模型从大量元素中进行选择，并通过自动相关性确定自适应地剪除冗余维度。事实上，我们的实验证实，较大的嵌入维度往往能提升性能，而自动相关性确定能有效剪除大多数嵌入维度（见第 7.2 节）。嵌入值是有序的而非二元的，这一特点可能有助于推断合适的长度尺度。由此得到命题 3 所示的搜索空间基数约简。

7 实验

我们在一系列具有挑战性的组合搜索空间和混合搜索空间优化问题上评估 BODi。我们将其与若干具有竞争力的基线方法进行比较，包括 CASMOPOLITAN、COMBO、CoCaBO、SMAC 和随机搜索。

实验设置。所有实验中，我们都使用期望改进作为采集函数。不过，请注意，我们的方法不受这一选择的限制，也可以采用任何其他采集函数，因此很容易将 BODi 扩展到多目标、多保真度和约束设置等场景。对于离散变量和连续变量，我们均采用带自动相关性确定（ARD）的 Mat´ern-5/2 核函数。在组合搜索空间中，我们使用爬山局部搜索来优化采集函数，方法与 CASMOPOLITAN [Wan 等人2021]所用的方法类似。对于所有实验，我们遵循算法 1（附录 G） $m = 128$，并 $m = 128$ 采用多样化随机方法来构建字典。我们在图 4c 的消融研究中考察了 $m = 128$ 这一选择。我们的代码基于常用的 GPyTorch [Gardner 等人， 2018] 和 BoTorch [Balandat 等人， 2020]库构建。所有基线方法均使用开源实现：CASMOPOLITAN [1]、COMBO [2]、CoCaBO [3], 和 SMAC [4]。

7.1 组合测试问题

LABS。低自相关二进制序列（LABS）问题的目标是找到一个二进制序列 $\{1, -1\}$ 其长度为 n，并使优值因子（*MF*）最大化：

$$\max_{\mathbf{x}\in\{1,-1\}^n} \mathrm{MF}(\mathbf{x}) = \frac{n^2}{E(\mathbf{x})},$$
$$E(\mathbf{x}) = \sum_{k=1}^{n-1}\left(\sum_{i=1}^{n-k} x_i x_{i+k}\right)^2$$

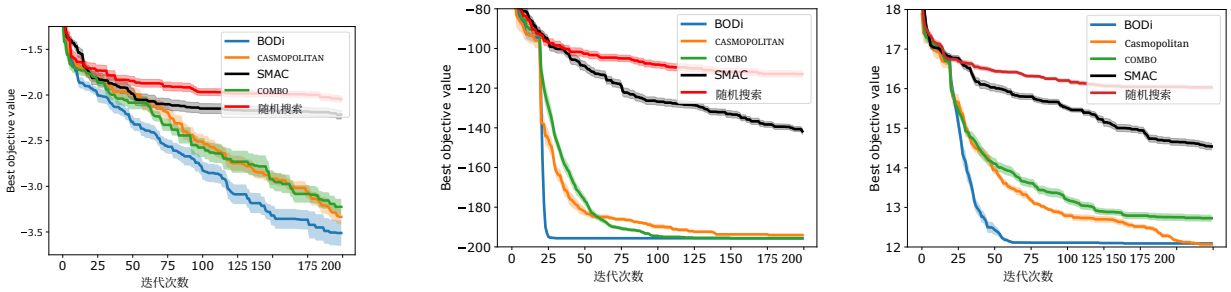
该问题在多个领域具有广泛的应用 [Bernasconi, 1987, Packebusch and Mertens, 2015],包括通信领域，在该领域中，它被用于高精度的行星际雷达测量，以测量时空曲率 [Shapiro et al., 1968]。我们在该问题的 50 维版本上评估所有方法。图 3a 绘制了负 MF，结果表明 BODi 找到的解显著优于基线方法。COMBO 和 CASMOPOLITAN 的表现虽不如 BODi，但找到的解优于 SMAC。随机方法的表现相当差，这表明对于具有挑战性的问题，采用模型引导的搜索技术十分重要（ LABS 的组合空间包含 $2^{50} \approx 1.2 \times 10^{15}$ 种构型）。需要注意的是，Packebusch and Mertens [2015]发布了 50 维 LABS 问题的优化器 $\mathbf{x}^{\text{opt}}$，其 $\mathrm{MF}(\mathbf{x}^{\text{opt}}) = 8.170$ 由分支定界算法计算得到，计算成本呈指数级增长。我们强调，这里的结果并不是要主张使用贝叶斯优化来求解这一特定的 LABS 问题，而是为了比较各贝叶斯优化算法在具有挑战性的组合优化任务上的表现；这些算法的设计目标是实现高效采样。

加权最大可满足性。该问题的目标是找到一个 60 维二进制向量，使满足的子句的权重总和最大化。我们采用最大可满足性竞赛 2018[6], 中的基准问题 frb-frb10-6-4.wcnf[5]，与 Oh 等人 [2019]和 Wan 等人 [2021]的做法类似。可满足性问题无处不在，经常出现在计算机科学的许多基础领域中 [Biere 等人， 2009]。图 3b 显示，尽管该组合搜索空间多达 $2^{60} \approx 1.2 \times 10^{18}$ 种可能配置，BODi 仍能迅速找到接近最优的解。BODi 在该问题上的出色表现，得益于在 HED 上训练的高斯过程具有优越的模型性能，详见第 7.2 节。

害虫防治。 该问题涉及控制一条由 25 个站点组成的链中的害虫传播。在每个站点，都可以从 5 个可能选项中进行分类选择，以使用一种杀虫剂；这些杀虫剂在成本和效果方面各不相同。由于配置总数为 $5^{25} \approx 3.0 \times 10^{17}$, 该问题颇具挑战性。从图 3c 中可以观察到，BODi 很快收敛到一个目标值

1https://github.com/xingchenwan/Casmopolitan
2https://github.com/QUVA-Lab/COMBO
3https://github.com/rubinxin/CoCaBO_code
4https://github.com/automl/SMAC3

5https://maxsat-evaluations.github.io/2018年/index.html
6http://sat2018.azurewebsites.net/competitions/



(a) LABS（50 个二元参数）

(b) 最大可满足性问题（60 个二元参数）

(c) 害虫防治（25 个类别参数，每个参数有 5 个可能取值）

图 3：我们在三个高维组合测试问题上比较了 BODi 与 CASMOPOLITAN、COMBO、SMAC 和随机搜索的表现。结果表明，BODi 的表现始终最佳，其次是 CASMOPOLITAN 和 COMBO。

约为 ≈ 12，并且在该问题上显著优于其他基线方法。

7.2 模型性能

为验证使用 HED 的高斯过程相较于分类重叠核（用于 CASMOPOLITAN，[Wan 等人，2021]）和扩散核（用于 COMBO，[Oh 等人，2019]）能否提供准确且校准良好的估计，我们考察了这些不同核函数在一个 60 维最大可满足性问题上的预测性能。我们生成了 50 个训练点和 50 个测试点，并将基于字典的核函数与高斯过程的测试预测结果，同重叠核和扩散核进行比较。测试集上的均值预测及其对应的 95% 预测区间如图 5 所示。

采用多样随机字典元素的 HED 能够准确地对未知黑盒函数建模，而重叠核函数和扩散核函数则无法准确预测测试数据。此外，我们还观察到，采用高斯随机字典的 HED（通过命题 1 的仿射表示计算）表现不佳。最后，尽管图 5 中使用的字典包含 $m = 128$ 个元素，但拟合后的 GP 模型中只有 4 个元素的长度尺度低于 10。这表明，自动相关性确定能够有效地剔除大多数字典元素，只保留少量元素，从而根据定理 4 得到更紧的遗憾界。

7.3 混合测试问题

混合 Ackley。我们考虑 [Wan 等人提出的标准 Ackley 问题的一个混合版本，其中包含 50 个二元变量和 3 个连续变量。可以看到，BODi 进展迅速，并逐渐接近全局最优值 0（图 4a）。除 CASMOPOLITAN 外，所有其他基线方法在此问题上的表现都很差。值得注意的是，经过子采样的二元小波字典在此问题上也表现出色，见附录图 7c。

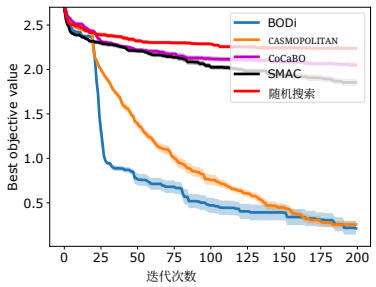
用于支持向量机训练的特征选择。在这个问题中，我们考虑联合进行特征选择和超参数优化，以在 UCI切片数据集 [Dua and Graff, 2019]上训练支持向量机（SVM）模型。我们对 50 个特征的纳入或排除进行优化，并进一步调整 SVM 的 C、 ϵ和 γ 超参数。目标是找到最优特征子集和连续超参数的取值，以使留出测试集上的均方根误差（RMSE）最小化。图 4b 显示，在这个真实世界问题上，BODi 的表现略优于 CASMOPOLITAN。

7.4 消融研究

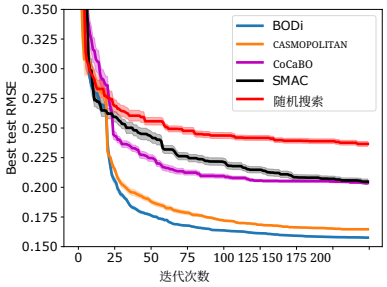
我们针对 BODi 对字典元素数量（字典大小）的敏感性进行了消融研究。我们考虑 50 维 LABS 问题。图 4c 中的结果表明，包含 $m = 128$ 个或$m = 256$ 个元素的字典表现最佳（尽管性能差异相对较小，至少对于较大的 m而言如此）。我们观察到，使用较小的字典（包含 $m = 16$个或 $m = 32$ 个元素）会导致性能较差。另一方面，使用大量元素会增加我们方法的运行时间，因此我们在所有实验中选择使用 $m = 128$ 个元素。

8 讨论

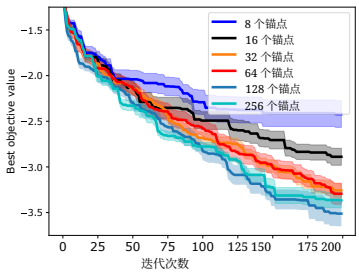
我们为高斯过程模型引入了一种新颖的字典核，适用于高维组合搜索空间（并且可以直接扩展到混合搜索空间）。虽然我们重点研究了将基于字典的建模方法用于贝叶斯优化，但我们的贡献所带来的影响远不止贝叶斯优化本身，对更广泛的基于核的方法同样具有意义。在贝叶斯优化的背景下，我们的字典核与采集函数的选择无关，可以轻松应用于多目标和多保真度优化等场景，也可以与信任



(a) 混合 Ackley（50 个二元参数，3 个连续参数)



(b) 支持向量机（50 个二元参数，3 个连续参数）



(c) LABS 消融实验（50 个二元参数）

图 4：（左、中）我们在两个同时包含离散参数和连续参数的高维问题上，将 BODi 与 CASMOPOLITAN、CoCaBO、SMAC 和随机搜索进行比较。对于 Ackley 问题，BODi 比 CASMOPOLITAN 收敛更快；对于支持向量机问题，BODi 的表现更好。（右）我们研究了 BODi 对字典大小（m）的敏感性，并观察到，只要字典中的元素数量不太少，性能就始终稳定。

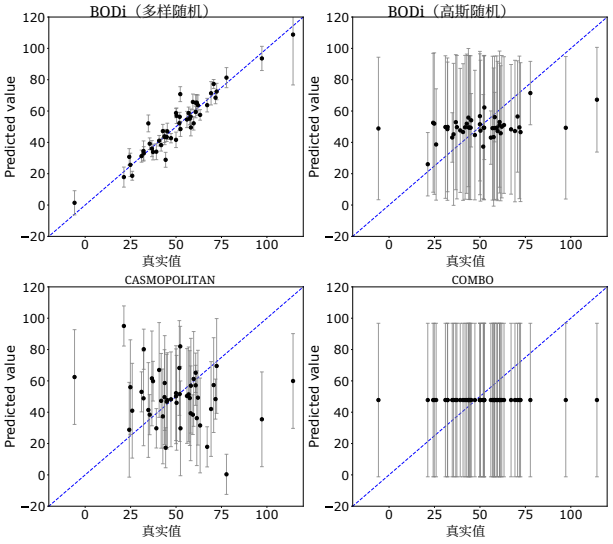


图 5：在最大可满足性问题上，BODi 使用多样随机字典（左上）、通过式（1）的仿射表示构造的高斯随机字典（右上）、Casmopolitan（左下）和 COMBO（右下）时的均值预测及相应的 95% 预测区间。我们使用 50 个训练点，并在 50 个测试点上进行预测。使用多样随机字典的 BODi 表现远优于使用高斯随机嵌入的情况，这验证了第 6 节中的理论结果。我们的核函数也优于 CASMOPOLITAN 使用的各向同性核函数和 COMBO 使用的扩散核。

区域优化等思想相结合。BODi 在一组多样化的问题上展现出强劲性能，并且优于多个实力强劲的基线方法，例如 CASMOPOLITAN 和 COMBO。

我们的工作存在一些局限，也提出了一些值得进一步探讨的有趣问题。尽管 BODi 不受采集函数选择的限制，但我们仅评估了它在单目标问题上的表现。此外，我们没有随机生成一组多样的字典元素，而是可以在模型拟合过程中优化字典，从而进一步改进基于字典的高斯过程模型。对于能够利用历史数据来帮助发现合适字典的情况，这种方法可能尤其有用。另一种思路是采用更符合贝叶斯优化目标的方式生成字典：贝叶斯优化的目标并非拟合全局准确的模型，而是找出全局最优解的位置。最后，BODi 还可能受益于近期提出的、用于混合搜索空间中高效采集函数优化的方法 [Daulton 等人， 2022]。

致谢。Aryan Deshwal 和 Jana Doppa 获得了美国国家科学基金会资助项目 IIS-1845922 和 OAC-1910213 的部分支持。

参考文献

M. Balandat、B. Karrer、D. R. Jiang、S. Daulton、B. Letham、A. G. Wilson 和 E. Bakshy。BoTorch：一种用于高效蒙特卡洛贝叶斯优化的框架。发表于《神经信息处理系统进展》第 *33* 卷：*2020*年神经信息处理系统大会，*NeurIPS 2020*，*2020*年*12*月*6*日至*12*日，线上，2020年。

R. Baptista 和 M. Poloczek。组合结构的贝叶斯优化。载于*ICML* 论文集，第 80 卷机器学习研究论文集，第 471–480 页。PMLR，2018 年。L. Baumert 、 S. W. Golomb 和 M. Hall Jr.。阶数 92 的哈达玛矩阵的发现。1962年。J. Bergstra、R. Bardenet、Y. Bengio 和 B. K´egl。超参数优化算法。载于神经信息处理系统进展 *24*：第 *25* 届年度

神经信息处理系统大会 *2011*。*2011*年*12*月*12*日至*14*日在西班牙格拉纳达举行的会议论文集，第 2546–2554 页，2011年。

J. Bernasconi。低自相关二进制序列：统计力学与构型空间分析。《物理学杂志》，48(4):559–567，1987年。

A. Biere、M. Heule 和 H. van Maaren。可满足性手册，第185卷。IOS出版社，2009年。

R. L. Clark、B. M. Connors、D. M. Stevenson、S. E. Hromada、J. J. Hamilton、D. Amador-Noguez 和 O. S. Venturelli。合成人类肠道微生物组组装与丁酸生产的设计。《自然通讯》，12(1):1–16，2021年。S. Daulton、X. Wan、D. Eriksson、M. Balandat、M. A. Osborne 和 E. Bakshy。通过概率重参数化在离散空间和混合空间上的贝叶斯优化。*arXiv*预印本 *arXiv:2210.10199*，2022年。A. Deshwal 和 J. R. Doppa。结合潜在空间与结构化核进行组合空间上的贝叶斯优化。载于神经信息处理系统进展（*NeurIPS*），第8185–8200页，2021年。A. Deshwal、S. Belakaria、J. R. Doppa 和 A. Fern。通过高成本评估优化离散空间：一种学习搜索框架。载于*AAAI*人工智能会议论文集，第34卷，第3773–3780页，2020年。A. Deshwal、S. Belakaria 和 J. R. Doppa。用于高效组合贝叶斯优化的 Mercer 特征。载于*AAAI*人工智能会议，2021a。A. Deshwal、S. Belakaria 和 J. R. Doppa。混合空间上的贝叶斯优化。载于*ICML*论文集，机器学习研究论文集第139卷，第2632–2643页。PMLR，2021b。D. Z. Djokovi´c、O. Golubitsky 和 I. S. Kotsireas。Hadamard矩阵和斜Hadamard矩阵的一些新阶数。《组合设计杂志》，22(6):270–277，2014年。J. R. Doppa。用于优化组合结构的自适应实验设计。载于第三十届国际人工智能联合会议（*IJCAI*）论文集，第4940–4945页，2021年。D. Dua 和 C. Graff。UCI机器学习库，2017年。网址：*http://archive.ics.uci.edu/ml*，7(1)，2019年。S. Eissman、D. Levy、R. Shu、S. Bartzsch 和 S. Ermon。贝叶斯优化与属性调整。载于第三十四届人工智能中的不确定性会议论文集，2018年。D. Eriksson 和 M. Jankowiak。利用稀疏轴对齐子空间进行高维贝叶斯优化。载于人工智能中的不确定性，第493–503页。PMLR，2021年。

D. Eriksson、M. Pearce、J. Gardner、R. D. Turner 和 M. Poloczek。通过局部贝叶斯优化实现可扩展的全局优化。神经信息处理系统进展，32，2019。P. I. Frazier。贝叶斯优化教程。*arXiv* 预印本，abs/1807.02811，2018。J. Gardner、C. Guo、K. Weinberger、R. Garnett 和 R. Grosse。发现并利用贝叶斯优化中的加性结构。载于人工智能与统计学，第1311–1319页。PMLR，2017。J. R. Gardner、G. Pleiss、K. Q. Weinberger、D. Bindel 和 A. G. Wilson。GPyTorch：利用 GPU 加速进行黑盒矩阵-矩阵高斯过程推断。载于神经信息处理系统进展第*31*卷：神经信息处理系统年度大会*2018*，*NeurIPS 2018*，*2018*年*12*月*3*日至*8*日，蒙特´利，加拿大，第7587–7597页，2018。R. Garnett、M. A. Osborne 和 P. Hennig。高斯过程线性嵌入的主动学习。*arXiv* 预印本 *arXiv:1310.6740*，2013。E. C. Garrido-Merch´an 和 D. Hern´andez-Lobato。在使用高斯过程进行贝叶斯优化时处理分类变量和整数值变量。*Neurocomputing*，380：20–35，2020。R. G´omez-Bombarelli、J. N. Wei、D. Duvenaud、J. M. Hern´andez-Lobato、B. S´anchez-Lengeling、D. Sheberla、J. Aguilera-Iparraguirre、T. D. Hirzel、R. P. Adams 和 A. Aspuru-Guzik。利用数据驱动的分子连续表示实现自动化学设计。*ACSCentral Science*，4(2)：268–276，2018。I. Guyon 和 A. Elisseeff。变量与特征选择简介。机器学习研究杂志，3(Mar)：1157–1182，2003。J. Hadamard。关于行列式的一个问题的解答。数学科学通报，2：240–246，1893年。A. Hedayat 和 W. D. Wallis。Hadamard 矩阵及其应用。统计年鉴，第1184–1238页，1978。E. Hellsten、A. Souza、J. Lenfers、R. Lacouture、O. Hsu、A. Ejjeh、F. Kjolstad、M. Steuwer、K. Olukotun 和 L. Nardi。Baco：快速且可移植的贝叶斯编译器优化框架。*arXiv* 预印本 *arXiv:2212.11142*，2022。K. J. Horadam。*Hadamard* 矩阵及其应用。普林斯顿大学出版社，2012。F. Hutter、H. H. Hoos 和 K. Leyton-Brown。用于通用算法配置的序贯模型优化。载于第*5*届学习与智能优化国际会议论文集，第507–523页。施普林格出版社，2011。ISBN 9783642255656。

H. Kajino。分子超图文法及其在分子优化中的应用。载于机器学习国际会议，第3183–3191页。PMLR，2019。

K. Kandasamy, J. Schneider, and B. P´oczos. 基于加性模型的高维贝叶斯优化与多臂老虎机. 收录于国际机器学习会议，第295–304页。PMLR，2015年。

J. Kim, S. Choi, and M. Cho. 利用随机映射函数对凸多胞体进行组合贝叶斯优化. 收录于人工智能中的不确定性，第1001–1011页。PMLR，2022年。J. Kirschner, M. Mutny, N. Hiller, R. Ischebeck, and A. Krause. 通过一维子空间实现高维空间中的自适应安全贝叶斯优化. 收录于第*36*届国际机器学习会议论文集，载于第97卷机器学习研究论文集，第3429–3438页。PMLR，2019年。K. G. Larsen and J. Nelson. Johnson–Lindenstrauss 引理的最优性. 收录于*2017*年 *IEEE* 第*58*届计算机科学基础年度研讨会（*FOCS*），第633–638页，2017年。doi：10.1109/FOCS.2017.64。B. Letham, R. Calandra, A. Rai, and E. Bakshy. 重新审视高维贝叶斯优化中的线性嵌入. 收录于神经信息处理系统第*33*届大会论文集：*2020*年神经信息处理系统年度大会（*NeurIPS 2020*），*2020*年*12*月*6*日至*12*日，线上，2020年。S. G. Mallat. 多分辨率信号分解理论：小波表示. *IEEE*模式分析与机器智能汇刊，11(7):674–693，1989年。N. Maus, H. T. Jones, J. S. Moore, M. J. Kusner, J. Bradshaw, and J. R. Gardner. 面向结构化输入的局部潜在空间贝叶斯优化. *CoRR*，abs/2201.11872，2022年。A. Nayebi, A. Munteanu, and M. Poloczek. 嵌入子空间中贝叶斯优化的框架. 收录于国际机器学习会议，第4752–4761页。PMLR，2019年。P. Notin, J. M. Hern´andez-Lobato, and Y. Gal. 利用解码器不确定性改进 VAE 潜在空间中的黑盒优化 . *arXiv* 预印本 *arXiv:2107.00096*，2021年。C. Oh, E. Gavves, and M. Welling. BOCK：使用圆柱核函数的贝叶斯优化. 收录于 J. G. Dy and A. Krause（编），第*35*届国际机器学习会议论文集（*ICML*），载于第80卷机器学习研究论文集，第3865–3874页。PMLR，2018年。C. Oh, J. M. Tomczak, E. Gavves, and M. Welling. 使用图笛卡尔

产品。在神经信息处理系统进展*32*：神经信息处理系统*2019*年年度会议，*NeurIPS 2019*，*2019*年*12*月*8*日至*14*日，加拿大不列颠哥伦比亚省温哥华，第2910–2920页，2019年。C. Oh、E. Gavves 和 M. Welling。基于频率调制核的混合变量贝叶斯优化。*arXiv*预印本，abs/ 2102，2021年。T. Packebusch 和 S. Mertens。低自相关二进制序列。物理学*A*杂志：数学与理论，*49*（*2016*）*165001*，2015年。doi：10.1088/ 1751-8113/49/16/165001。L. Papenmeier、L. Nardi 和 M. Poloczek。随着学习扩大范围：嵌套子空间中的自适应贝叶斯优化。载于神经信息处理系统进展（*NeurIPS*），2022年。C. E. Rasmussen。机器学习中的高斯过程。载于机器学习高级讲座：*2003*年机器学习暑期学校，澳大利亚堪培拉，*2003*年*2*月*2*日至*14*日，*T¨*德国图宾根，*2003*年*8*月*4*日至*16*日，修订版讲义，2004年。B. X. Ru、A. S. Alvi、V. Nguyen、M. A. Osborne 和 S. J. Roberts。针对多个连续型和类别型输入的贝叶斯优化。载于*ICML*论文集，机器学习研究论文集第119卷，第8276–8285页。PMLR，2020年。W. Rudin。实分析与复分析，麦格劳-希尔。公司，1974年。I. I. Shapiro、G. H. Pettengill、M. E. Ash、M. L. Stone、W. B. Smith、R. P. Ingalls 和 R. A. Brockelman。广义相对论第四次检验：初步结果。物理评论快报，20(22):1265，1968年。N. Srinivas、A. Krause、S. M. Kakade 和 M. W. Seeger。多臂老虎机设定中的高斯过程优化：无遗憾与实验设计。载于*ICML*论文集，第1015–1022页。Omnipress，2010年。M. D. Swanson 和 A. H. Tewfik。二值图像的二进制小波分解。*IEEE*图像处理汇刊，5(12):1637–1650，1996年。A. Tripp、E. Daxberger 和 J. M. Hern´andez-Lobato。通过加权再训练在深度生成模型的潜在空间中进行样本高效优化。神经信息处理系统进展，33，2020年。J. A. Tropp。贪婪有益：稀疏逼近的算法结果。*IEEE*信息论汇刊，50(10):2231–2242，2004年。X. Wan、V. Nguyen、H. Ha、B. X. Ru、C. Lu 和 M. A. Osborne。着眼全局，立足局部：针对高维类别型和混合搜索空间的贝叶斯优化。载于*ICML*论文集，机器学习研究论文集第139卷，第10663–10674页。PMLR，2021年。

Z. Wang、F. Hutter、M. Zoghi、D. Matheson 和 N. De Feitas。通过随机嵌入在十亿维空间中进行贝叶斯优化。人工智能研究杂志，55:361–387，2016年。X. Zhang、Z. Chang、Y. Li、H. Wu、J. Tan、F. Li 和 B. Cui。利用超参数优化促进数据库调优：全面的实验评估。*arXiv*预印本 *arXiv:2110.12654*，2021年。

通过基于字典的嵌入进行高维组合空间的贝叶斯优化补充材料

汉明嵌入的仿射表示

我们的第一个命题表明， $\{0,1\}^d$ 中向量的汉明嵌入等价于原始二进制向量的 $\{-1,1\}$编码的仿射变换。

命题 1 （仿射表示）。 设 $\mathbf{A} \in \{0,1\}^{n\times d}$， $\mathbf{z} \in \{0,1\}^d$。则

$$2\boldsymbol{\phi}_{\mathbf{A}}(\mathbf{z}) = d\mathbf{1}_n - \bar{\mathbf{A}}\bar{\mathbf{z}}, \tag{2}$$

¯其中 $A_{ij} = 2A_{ij} - 1$ 且 $\bar{z}_i = 2z_i - 1 \in \{-1,1\}$。

证明。 令$\mathbf{a}_i$ 为$\mathbf{A}$中的第 i 列，并令 z_i 为$\mathbf{z}$中的第 i 个元素。于是

$$\begin{aligned}
\boldsymbol{\phi}_{\mathbf{A}}(\mathbf{x}) &= \sum_i^d (\neg\mathbf{a}_i z_i + \mathbf{a}_i \neg z_i) \\
&= \sum_i^d ([\mathbf{1}_n - \mathbf{a}_i] z_i + \mathbf{a}_i [1 - z_i]) \\
&= \sum_i^d (\mathbf{1}_n z_i - 2\mathbf{a}_i z_i + \mathbf{a}_i) \\
&= \mathbf{1}_n(\mathbf{1}_d^\top \mathbf{z}) - \mathbf{A}(2\mathbf{z} - 1) \\
&= [(\mathbf{1}_n \mathbf{1}_d^\top)(2\mathbf{z} - 1) + d\mathbf{1}_n]/2 - \mathbf{A}(2\mathbf{z} - 1) \\
&= [d\mathbf{1}_n - (2\mathbf{A} - \mathbf{1}_{n,d})(2\mathbf{z} - \mathbf{1}_d)]/2 \\
&= (d\mathbf{1}_n - \bar{\mathbf{A}}\bar{\mathbf{z}})/2.
\end{aligned}$$

将等式两边乘以二，即可完成证明。 □

将仿射表示代入嵌入距离公式，得到

$$\begin{aligned}
2\|\boldsymbol{\phi}_{\mathbf{A}}(\mathbf{z}) - \boldsymbol{\phi}_{\mathbf{A}}(\mathbf{z}')\| &= \|(d\mathbf{1}_n - \bar{\mathbf{A}}\bar{\mathbf{z}}) - (d\mathbf{1}_n - \bar{\mathbf{A}}\bar{\mathbf{x}}')\| \\
&= \|\bar{\mathbf{A}}\bar{\mathbf{z}} - \bar{\mathbf{A}}\bar{\mathbf{z}}'\| \\
&= \|\bar{\mathbf{A}}\bar{\mathbf{r}}\|,
\end{aligned}$$

其中 $\bar{\mathbf{r}} = \bar{\mathbf{z}} - \bar{\mathbf{z}}'$。也就是说，距离计算仅依赖于差向量$\mathbf{r}$ 对中心化输入向量的$\{-1,1\}$编码所作的线性 投影。进一步而言，如果选择 C 节中的小波字典，那么该嵌入在一个常数平移的意义下是经子采样的哈达玛变换，我们可以通过快速哈达玛变换在 $d\log\ d$ 时间内实现它。

仿射表示的另一个推论是，汉明距离 h 可以先写成平移后输入 ¯$\mathbf{z}$之间的欧几里得距离。此外，我们可以利用$\|\mathbf{z}\|_2^2 = d$ 这一事实，写出

$$2h(\mathbf{z},\mathbf{z}') = d - \bar{\mathbf{z}}^\top\bar{\mathbf{z}}' = (\|\bar{\mathbf{z}}\|_2^2 + \|\bar{\mathbf{z}}'\|_2^2)/2 - \bar{\mathbf{z}}^\top\bar{\mathbf{z}}' = \|\bar{\mathbf{z}} - \bar{\mathbf{z}}'\|_2^2/2.$$

因此，汉明距离负值的指数可以视为径向基函数核：

$$\exp(-2h(\mathbf{z},\mathbf{z}')) = \exp(-\|\bar{\mathbf{z}} - \bar{\mathbf{z}}'\|_2^2/2). \tag{3}$$

B 搜索空间基数缩减

本节推导嵌入后输入空间的基数上界 $\phi_{\mathbf{A}}(\mathbf{z})$。使用字典核等价于对变换后的搜索空间 $\mathcal{S} = \{\phi_{\mathbf{A}}(\mathbf{z}) \mid \mathbf{z} \in \{0,1\}^d\}$应用典范核。因此，有限搜索空间的一般收敛界和遗憾界均适用。然而，尽管一般的线性嵌入通常会降低搜索空间的维度，却不一定会降低其基数 $|\mathcal{S}|$；而基数是有限搜索空间贝叶斯优化遗憾界中的一个关键量。事实上，接下来的结果表明，即使采用一维高斯随机投影，完整的基数仍会保留。

命题 2。 定义 $\mathcal{S}_{\mathbf{a}} = \{\mathbf{a}^\top \mathbf{z} \mid \mathbf{z} \in \{\pm 1\}^d\}$，并令 $\mathbf{a} \sim \mathcal{N}(\mathbf{0},\ \mathbf{I}_d)$。则 $|\mathcal{S}_{\mathbf{a}}| = 2^d$ 几乎必然成立。

证明。给定 $\mathbf{z},\ \mathbf{z}' \in \{-1,1\}^d$，设 $\mathbf{z} \neq \mathbf{z}'$ 且 $\mathbf{a}^\top \mathbf{z} = \mathbf{a}^\top \mathbf{z}'$。因此，$\mathbf{a}^\top(\mathbf{z} - \mathbf{z}') = 0$。由于 $\mathbf{z} - \mathbf{z}' \neq 0$，这只有在 $\mathbf{a} \perp (\mathbf{z} - \mathbf{z}')$ 时才可能成立。但 $\{\mathbf{a} \mid \mathbf{a} \perp (\mathbf{z} - \mathbf{z}')\}$ 是 $(d-1)$ 维的，因此在 d 维空间中的高斯测度下构成零测集 [Rudin, 1974]。因此，$\mathbf{a}^\top(\mathbf{z} - \mathbf{z}') \neq 0$ 几乎必然成立。由于集合 $\{-1,1\}^d$ 具有有限基数 2^d，并且根据任意概率测度的次可加性 μ，

$$\mu\left(\bigcup_{\mathbf{z},\mathbf{z}' \in \{-1,1\}^d} \{\mathbf{a} \mid \mathbf{a} \perp (\mathbf{z} - \mathbf{z}') = 0\}\right) \leq \sum_{\mathbf{z},\mathbf{z}' \in \{-1,1\}^d} \mu\left(\{\mathbf{a} \mid \mathbf{a} \perp (\mathbf{z} - \mathbf{z}') = 0\}\right) = 0.$$

因此，几乎必然地，所有不同的 $\mathbf{z} \in \{-1,1\}^d$ 都映射到不同的值 $\mathbf{a}^\top \mathbf{z}$，所以 $|\mathcal{S}_{\mathbf{a}}| = |\{-1,1\}|^d = 2^d$。 □

以下命题阐明了嵌入搜索空间的隐含基数如何随嵌入维度的数量 n、输入维度 d以及字典行变异性的一种度量而变化。

命题 3 (嵌入 基数)。 设 $\mathbf{A} \in \{0,1\}^{m \times d}$。那么嵌入搜索空间 $\mathcal{S}_{\mathbf{A}}$ 的基数可以被上界限定为

$$|\mathcal{S}_{\mathbf{A}}| \leq \left[(\mu_{\mathbf{A}} + 1)(d + 1 - \mu_{\mathbf{A}})\right]^{\lfloor m/2 \rfloor} (d+1)^{m \bmod 2}$$

其中 $\mu_{\mathbf{A}} = \max_{i,j} \max(h(\mathbf{a}_i, \mathbf{a}_j),\ h(\neg\mathbf{a}_i, \mathbf{a}_j))$，且 h 是汉明距离。

证明。 首先，我们考虑一个锚点。设 $d \in \mathbb{N}$，且 $\mathbf{a} \in \mathcal{B}^d$。那么对于任意 $\mathbf{z} \in \mathcal{B}^d$, $\phi(\mathbf{a},\mathbf{z}) \in \mathbb{N}$ 以及

$$0 \leq \phi_{\mathbf{a}}(\mathbf{z}) = h(\mathbf{a}, \mathbf{z}) = \sum_i \delta(a_i, z_i) \leq d,$$

因此 $\phi_{\mathbf{a}}(\mathbf{z}) \in [d]$ 且 $|\mathcal{S}| = d + 1$。朴素地将其推广到 n 个维度，会得到 $|\mathcal{S}| \leq (d+1)^n$。然而，实际基数要小得多，因为与一个锚点具有某些相同元素，会限制相应的维度，使其与另一个锚点的对应维度相同。下一段将对这一直觉作出精确说明。

接下来，我们考虑两个锚点。设 $d \in \mathbb{N}$，且 $\mathbf{a}_1, \mathbf{a}_2 \in \mathcal{B}^d$。那么对于任意 $\mathbf{z} \in \mathcal{B}^d$，假设 $\mathbf{A} = [\mathbf{a}_1,\mathbf{a}_2],$，并令

$$s = \{i \in [d] \mid [\mathbf{a}_1]_i = [\mathbf{a}_2]_i\}$$

表示锚点取值相同的指标集合，且 $\neg s = [d] \backslash s$, $|s| = k$。那么我们可以将嵌入表示为

$$\begin{aligned}
\phi_{\mathbf{A}}(\mathbf{z}) &= \phi_{\mathbf{A}_s}(\mathbf{z}_s) + \phi_{\mathbf{A}_{\neg s}}(\mathbf{z}_{\neg s}) \\
&= [h(\mathbf{a}_{1,s}, \mathbf{z}_s), h(\mathbf{a}_{2,s}, \mathbf{z}_s)] + [h(\mathbf{a}_{1,\neg s}, \mathbf{z}_{\neg s}), h(\mathbf{a}_{2,\neg s}, \mathbf{z}_{\neg s})] \\
&= [h(\mathbf{a}_{1,s}, \mathbf{z}_s), h(\mathbf{a}_{1,s}, \mathbf{z}_s)] + [h(\mathbf{a}_{1,\neg s}, \mathbf{z}_{\neg s}), h(\neg\mathbf{a}_{1,\neg s}, \mathbf{z}_{\neg s})] \\
&= [z_s, z_s] + [h(\mathbf{a}_{1,\neg s}, \mathbf{z}_{\neg s}), (d - h(\mathbf{a}_{1,\neg s}, \mathbf{z}_{\neg s}))] \\
&= [z_s, z_s] + [z_{\neg s}, (d - z_{\neg s})],
\end{aligned}$$

其中 $z_s = h(\mathbf{a}_{1,\ s},\ \mathbf{z}_s)$。现在，$n_s = h(\mathbf{a}_s,\ \mathbf{z}_s) \in [k]$ 以及 $n_{\neg s} \in [d-k]$。嵌入空间的基数恰好为 $(k+1)(d+1-k)$，因为有一部分 $d-k$ 变量在两个维度中始终取相同的值，而其余 k 个变量与前者线性无关。对 k 的基数求导：

$$\frac{d}{dk}(k+1)(d+1-k) = d - 2k \leq 0 \qquad \text{for} \qquad \lceil d/2 \rceil \leq k \leq d,$$

我们可以看到，基数是关于 $k = \lceil d/2 \rceil$ 的偶对称函数，并在此处达到最大值。这启发我们定义相干性类量 $\mu_{\mathbf{A}}$，其取值与上述基数公式单调相关，并满足 $\lceil d/2 \rceil \leq \mu_{\mathbf{A}} \leq d$。此外，注意对于两个锚点，$\mu_{\mathbf{A}} = \max(h(\neg \mathbf{a}_1, \mathbf{a}_2)\ h(\mathbf{a}_1, \mathbf{a}_2)) = \max(k,\ d-k)$。对于 m 行，$\mu_{\mathbf{A}}$ 是所有行及其取反之间任意成对相似度的上界。因此，我们可以将上述界应用于 $\lfloor m/2 \rfloor$ 对，并且如果 m 为奇数，则剩余维度最多再有 $(d+1)$ 个值。 □

现在，我们已准备好将对嵌入搜索空间 $\mathcal{S}_{\mathbf{A}}$ 基数的分析与 Srinivas 等人 [2010] 的一般结果结合起来，从而为 BODi 得到更优的遗憾界。我们将使用 $\mathcal{O}^*$ 表示 $\mathcal{O}$，其中省略了对数因子。

定理 4。设$\mathbf{A}$ 具有 m 行，$\delta \in (0、1)$，且 $\beta_t = 2\log(|\mathcal{S}_{\mathbf{A}}| t^2 \pi^2 / 6\delta)$。则对于从核函数为 $k_{BODi}(\mathbf{z}、\mathbf{z}') = k_{base}(\phi_{\mathbf{A}}(\mathbf{z})\ \phi_{\mathbf{A}}(\mathbf{z}'))$的零均值高斯过程进行 f 采样时，运行置信上界所对应的累积遗憾以概率 $1-\delta$被上界 $\mathcal{O}^*(\sqrt{T\gamma_T m})$，其中 γ_T是 k_{base} 的最大信息增益。

证明。 BODi 等价于在变换后的搜索空间 $\mathcal{S}_{\mathbf{A}} = \left\{\phi_{\mathbf{A}}(\mathbf{z}) \mid \mathbf{z} \in \{0,1\}^d\right\}$上，使用 k_{base} 核函数运行标准贝叶斯优化。由于 $\mathcal{S}_{\mathbf{A}}$ 是有限的，Srinivas 等人的定理 1 [2010] 适用，其中搜索空间为 $\mathcal{S}_{\mathbf{A}}$，基核函数 k_{base}的信息增益为 γ_T，从而得到遗憾界 $\mathcal{O}^*(\sqrt{T\gamma_T \log(|\mathcal{S}_{\mathbf{A}}|)})$。应用命题 3 的基数界，可得 $\mathcal{O}^*(\log(|\mathcal{S}_{\mathbf{A}}|)) = \mathcal{O}^*(\log([(\mu_{\mathbf{A}}+1)(d+1-\mu_{\mathbf{A}})]^{\lfloor m/2 \rfloor})) = \mathcal{O}^*(m)$。将此基数界代入通用渐近界，即完成证明。 □

通过二元小波构造 C 词典的方法

本节介绍一种基于二进小波变换、适用于二元空间 $\mathcal{Z}$=$\{0,1\}^d$ 的随机词典构造方法。从整体上看，该方法包含两个关键步骤。首先，我们采用确定性递归过程，在二进制结构上构造一个基向量池。其次，我们从 k 个多样化向量中随机选取一个子集作为词典 $\mathbf{A}$。下面将详细说明这两个步骤。

用于二进小波设计的递归算法。代理模型的有效性在很大程度上取决于用于嵌入离散输入的字典。我们将字典矩阵 $\mathbf{A}^{[k \times d]}$ 定义为二元空间 $\{0,1\}^d$ 上的一组经过子采样（大小为 k）的基向量，这些基向量的序列度各不相同。序列定义为从 1 变为 0 或从 0 变为 1 的次数（类似于傅里叶变换中的频率概念）。

多分辨率小波 [（Mallat， 1989] ）是研究不同尺度下实值信号的有效且广为人知的方法，它通过对数据应用一组正交变换来实现这一点。具体而言，二元小波变换 [（Swanson 和 Tewfik，1996] ）使我们能够研究定义在二元空间上的数据（具体来说，$\{0,1\}^d$ 采用模 *2*运算）。因此，它们是构建基向量池的自然选择。

我们通过从确定性二进小波变换矩阵$\mathbf{B}_d$中随机抽样，构造随机字典$\mathbf{A}$。该矩阵由递归过程生成，具体如 [Swanson 和 Tewfik 1996] 所述（此类矩阵曾用于图像压缩）。这一过程的核心思想是递归生成二进制矩阵，并按序列递增的顺序排列其向量。算法 3 给出了这一递归方法的伪代码。

给定$\mathbf{B}_d$，通过对$\mathbf{B}_d$ 的行向量进行下采样来构造字典$\mathbf{A}$，即$\mathbf{A} = \mathbf{P}\mathbf{B}_d$，其中$\mathbf{P}$ 以均匀分布随机抽样m 向量。使用$\mathbf{P}$ 进行随机抽样，选出的向量会分布在较广的序列度范围内；相比之下，另一种做法是从$\mathbf{B}_d$ 中选取前-m 行，这会将所选向量限制在较窄的序列度范围内。我们的实验表明，随机字典优于选取前-m 行的方案。

备注。 根据命题 1，这种字典选择等价于用于在连续输入空间 中构造低维嵌入的子采样随机哈达玛变换（ SRHT），其中嵌入是对哈达玛变换进行子采样得到的投影，即$\hat{}\mathbf{x} = \mathbf{P}\mathbf{H_n}\mathbf{D}\mathbf{x}$ 其中$\mathbf{H}_n$ 是阶数为 n 的哈达玛矩阵，$\mathbf{D}$ 是对角线上元素从 $\{1,-1\}$ 中随机选取的对角矩阵，而$\mathbf{P}$ 的定义与上文类似。重要的是，该字典还使命题 3 中引入的相干性类度量 $\mu_{\mathbf{A}}$ 最小化。

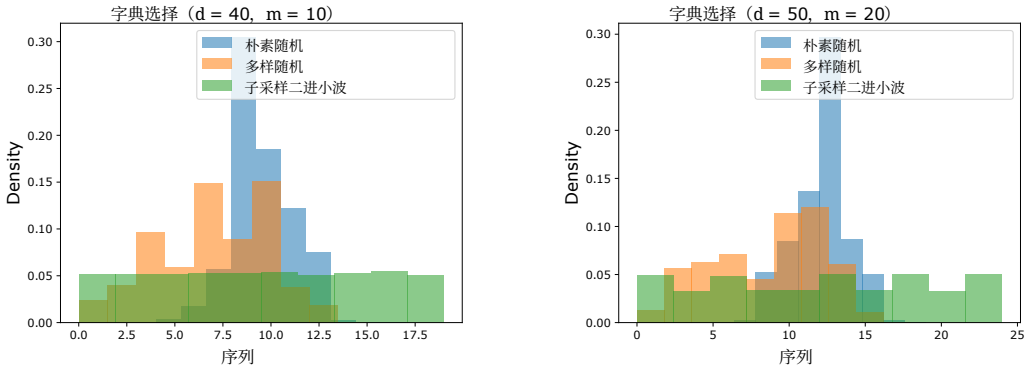
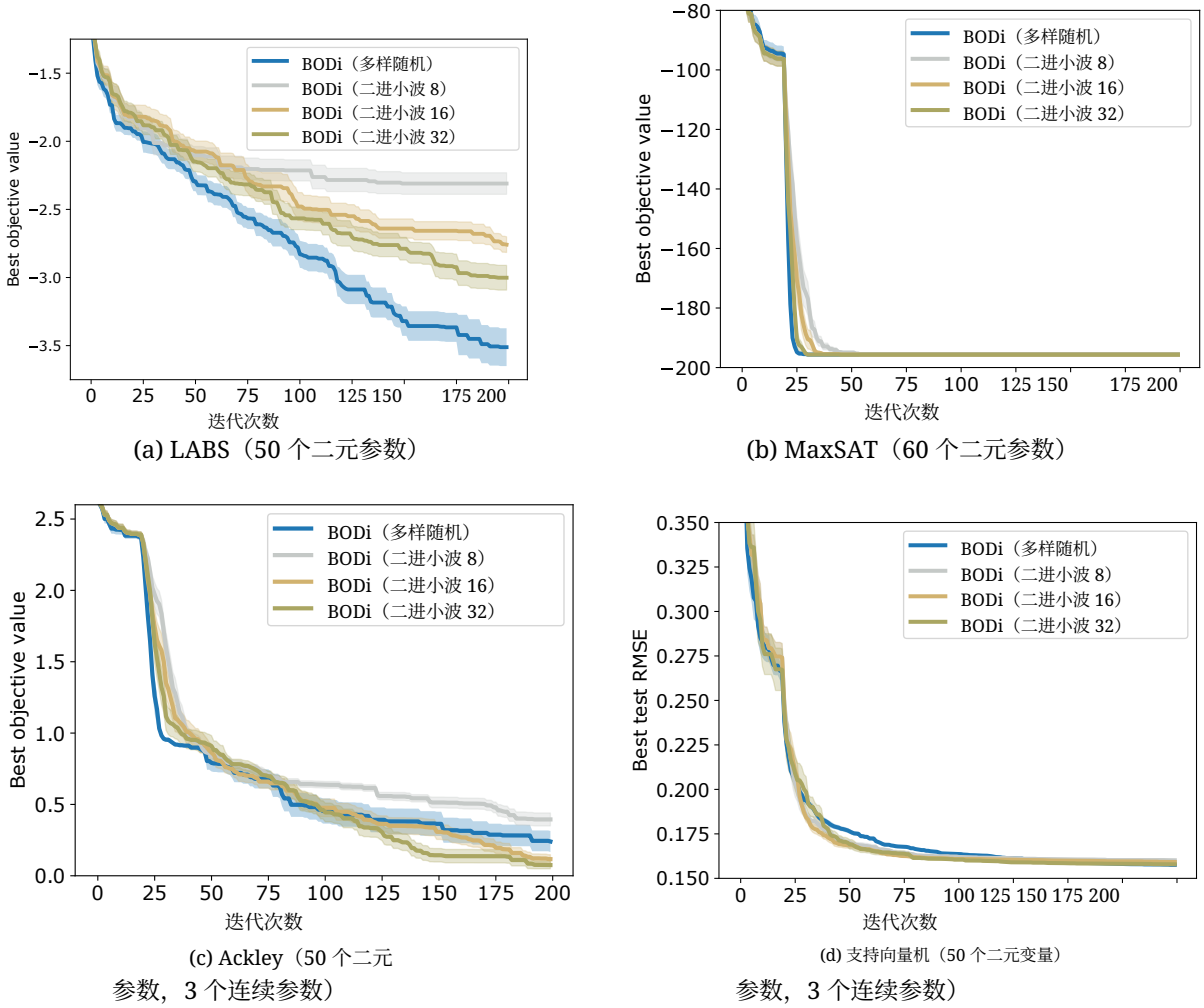


图6：我们随机生成了大量（1000 个）字典。图中展示了两种字典选择（二进制小波）和（多样参数）在序列特征刻画方面的相似性。序列这一概念使我们能够通过实证观察到这两种更优字典选择之间的相似性。

算法 3 二进制小波（n）变换需要：输入维度 n1：如果 $n == 2$：返回 $\begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$2：如果 $n == 4$：返回 $\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix}$3：$B_{n-4}$= 二进制小波（$n$-4）4：计算左上角的 $n-2 \times n-2$ 矩阵 $\Gamma = \begin{bmatrix} 2 \times \mathbf{1}_{[2,2]} - 2 & \neg B_{n-4} \\ \mathbf{1}_{[n-4,2]} & \neg B_{n-4} \end{bmatrix}$5：设置左下角块 $\Delta^T \leftarrow \begin{bmatrix} 1 & 0 & 1 & \cdots \\ 1 & 0 & 1 & \cdots \end{bmatrix}$6：设置右下角块 $\Lambda \leftarrow \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$7：返回 $B_n = \begin{bmatrix} \Gamma & \Delta \\ \Delta^T & \Lambda \end{bmatrix}$

D 定位 用于优化组合空间上采集函数的搜索 es

在每次优化采集函数的迭代中，我们首先生成一组初始输入，作为在组合输入上进行局部搜索的起点。这些初始输入是从均匀生成的随机输入和喷洒输入（基于汉明距离、邻近贝叶斯优化运行中当前最佳未覆盖输入的输入）组成的集合中，挑选排名靠前的候选项构造而成。对于每个起始输入，我们运行贪心爬山搜索，每次移动到采集函数值最高的汉明距离为 1 的邻居，直到搜索收敛或达到 n_{ls} 次迭代上限。然后，从所有局部搜索轨迹中选出最佳候选项，作为下一个待评估的输入。



(a) LABS（50 个二元参数）

(b) MaxSAT（60 个二元参数）

(c) Ackley（50 个二元参数，3 个连续参数）

(d) 支持向量机（50 个二元变量参数，3 个连续参数）

图 7：比较 BODi 的两种字典构造选择（即多样随机和二进制小波）的结果。总体而言，我们发现二进小波设计表现尚可，但综合考虑所有基准测试，多样随机是更稳健的选择。此外，与仅适用于二元参数的二进制小波构造不同，多样随机也可用于类别参数。

E 运行时间

图 8 展示了 BODi、COMBO 和 CASMOPOLITAN 的运行时间对比。我们展示了在具有 60 个二元参数的最大可满足性问题上，拟合模型并生成新候选项所需的平均时间。BODi 使用默认的 128 个锚点，这一设置也用于所有实验。我们发现，BODi 和 CASMOPOLITAN 的速度显著快于 COMBO，平均每次贝叶斯优化迭代耗时不到 10 秒。

F 社会影响力

贝叶斯优化是广泛应用于各类黑盒优化问题的一种常用方法，其中包括自动化机器学习（自动机器学习）。我们提出的 BODi 方法的主要优势在于，能够提升贝叶斯优化在组合型及混合搜索空间上的优化性能——在自动机器学习的情境下，这意味着能够找到更优的模型，或以更少的计算资源找到同样出色的模型。我们认为，除对底层应用可能被滥用的更普遍担忧之外，此类改进带来的额外风险极小。

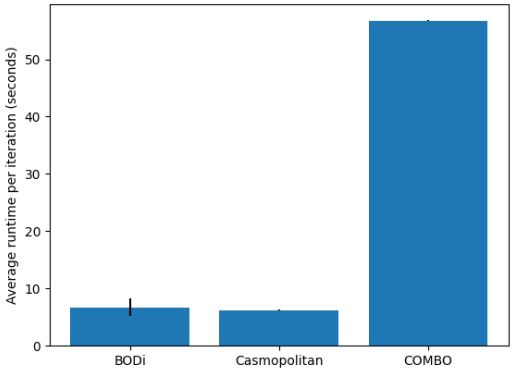


图 8：BODi 与 CASMOPOLITAN 和 COMBO 每次迭代的平均运行时间（秒）。

G 二元变量和类别变量字典的字典构造

算法 4 给出了在二元输入空间 $\{0,1\}^d$上构建字典的伪代码。其核心思想是：通过生成由伯努利分布的不同偏置参数（θ）决定的二元向量，来增加所构建字典的多样性（不同于朴素随机方法，其中 θ 始终为 $1/2$）。该算法以如下方式推广到大小可变的类别型输入（算法 5）：对于字典 $\mathbf{A}$ 的 m 个元素中的每一个，我们首先从 τ_{max}-单纯形 $\Delta^{\tau_{max}}$中采样一个权重向量 θ，其中 $\tau_{max} = \max_j \tau_j$。随后，对于每个变量 v_j，我们从 θ 中采样τ_j 个元素，并将其用作分类分布的权重向量，再从该分布中抽取字典元素的第 j 个维度。

算法 4 具有不同稀疏行的二元输入空间字典设计 $\{0,1\}^d$

要求：字典大小 m

1：字典 $\mathbf{A}$ ←为空
2：**对于** i=1, 2, . . . , m **do**
3：　一个$_i$ ←空
4：　采样伯努利参数 $\theta \sim$ 均匀分布(0，1)
5：　**对于** j=1, 2, . . . , d **执行**
6：　　采样二进制数 $a \sim$ 伯努利(θ)
7：　　一个$_i \leftarrow$ 一个$_i \cup a$
8：　**结束循环**
9：　将 一个$_i$ 添加到字典：$\mathbf{A} \leftarrow \mathbf{A} \cup$一个$_i$
10：**结束循环**
11：**返回**大小为 $m \times d$ 的字典 $\mathbf{A}$

算法 5 示意图

我们通过一个简单示例来说明算法 5 中的描述：输入的每个维度都有相同数量的候选选项。假设我们要为一个包含 10 个变量的输入空间（即 10 维输入）构造一个字典向量，其中每个维度都可以取 4 个值（即 $\tau_1 = \tau_2 \ldots = \tau_d = 4$）。对于每个输入维度（第 6 行的 for 循环），我们从由权重 θ_j 参数化的分类分布中采样一个值。在朴素随机方法中，θ_j 为 [1⁄4、1⁄4、1⁄4、1⁄4]。相比之下，算法 5 通过在第 5 行从单纯形中采样，使该向量 θ_j 具有多样性。变量 τ_{max}（第 2 行）以及第 7 行对 θ 的重新采样，使我们能够将算法推广到每个输入维度可取不同数量值的情况。

算法 5：通过多样化参数为包含类别变量的离散空间设计字典 输入：候选集 $C(v_1), \ldots, C(v_d)$, 字典大小 m 输出：大小为 $m \times d$ 的字典 A 1：字典 A $\leftarrow$ 为空 2：$\tau_{max} \leftarrow$ 最大值 $j\ \tau_j$ 3：对于 i=1, 2, . . . , m 执行 4：$\mathbf{a}_i \leftarrow$为空 5：采样 $\theta \sim \Delta^{\tau_{max}}$ 6：对于 j=1, 2, . . . , d 执行 7：从 θ中$\theta_j \leftarrow$ 采样（不放回）τ_j 个元素 8：$\theta_j \leftarrow \theta_j / \|\theta_j\|_1$ （归一化以得到分布） 9：从 $C(v_j)$ 中按概率 θ_j采样$a \leftarrow$ 10：$\mathbf{a}_i \leftarrow \mathbf{a}_i \cup a$ 11：结束循环 12：将 $\mathbf{a}_i$ 添加到字典：A $\leftarrow$ A $\cup \mathbf{a}_i$ 13：结束循环