软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, [doi: 10.13328/j.cnki.jos.005607]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

贝叶斯优化方法和应用综述*

崔佳旭[1,2], 杨 博[1,2]

[1](吉林大学 符号计算与知识工程教育部重点实验室,吉林 长春 130012)

[2](吉林大学 计算机科学与技术学院,吉林 长春 130012)

通讯作者: 杨博, E-mail: ybo@jlu.edu.cn

摘 要: 设计类问题在科学研究和工业领域无处不在.作为一种十分有效的全局优化算法, 近年来, 贝叶斯优化方法在设计类问题上被广泛应用. 通过设计恰当的概率代理模型和采集函数,贝叶斯优化框架只需经过少数次目标函数评估即可获得理想解, 非常适用于求解目标函数表达式未知、非凸、多峰和评估代价高昂的复杂优化问题.从方法论和应用领域两方面深入分析、讨论和展望了贝叶斯优化的研究现状、面临的问题和应用领域, 期望为相关领域的研究者提供有益的借鉴和参考.

关键词: 贝叶斯优化;全局优化算法;概率代理模型;采集函数;黑箱

中图法分类号: TP311

中文引用格式: 崔佳旭,杨博.贝叶斯优化方法和应用综述.软件学报. http://www.jos.org.cn/1000-9825/5607.htm

英文引用格式: CUI JX, YANG B. Survey on Bayesian Optimization Methodology and Applications. Ruan Jian Xue Bao/Journal of Software, (in Chinese). http://www.jos.org.cn/1000-9825/5607.htm

Survey on Bayesian Optimization Methodology and Applications

CUI Jia-Xu[1,2], YANG Bo[1,2]

[1](Key Laboratory of Symbolic Computation and Knowledge Engineering for the Ministry of Education, Jilin University, Changchun 130012, China)

[2](College of Computer Science and Technology, Jilin University, Changchun 130012, China)

Abstract: Designing problems are ubiquitous in the science research and industry fields. In recent years, Bayesian Optimization, which acts as a very effective global optimization algorithm, has been widely applied in designing problems. By structuring the probabilistic surrogate model and the acquisition function appropriately, Bayesian Optimization framework can guarantee to get the optimal solution under a few numbers of function evaluations, so it is very suitable to solve the extremely complex optimization problems that their objective functions could not be expressed, or they are non-convex, multimodal and evaluated expensively. This paper deeply analyses Bayesian Optimization from methodology and application fields, and then discusses its research status and the problems that we may face in future researches. This work is hopefully beneficial to the researchers from the related communities.

Key words: Bayesian Optimization; global optimization algorithm; probabilistic surrogate model; acquisition function; black-box

1 引言

设计类问题在科学研究和工业设计等领域无处不在.例如:编程人员通过选择恰当的算法来优化系统性能;环境学家通过设计传感器部署位置来监控环境状况;化学家通过设计实验来获取新的物质;制药厂商通过设计

* 基金项目: 国家自然科学基金(61373053, 61572226); 吉林省重点科技研发项目(20180201067GX,20180201044GX)

Foundation item: National Natural Science Foundation of China (61373053, 61572226); Jilin Province Key Scientific and Technological Research and Development project under grants 20180201067GX and 20180201044GX.

收稿时间: 2017-06-12; 修改时间: 2018-04-02; 采用时间: 2018-05-17; jos 在线出版时间: 2018-06-07

CNKI 网络优先出版: 2018-06-07 14:53:49, http://kns.cnki.net/kcms/detail/11.2560.TP.20180607.1453.008.html

新型药物来抵抗疾病;食品厂商通过设计新的食谱来生产优质食品等等.通常将这些设计问题考虑成如下最优化问题加以求解[I]:

$$\boldsymbol{x}^* = \arg\min_{\boldsymbol{x}\in\mathcal{X}\subseteq\mathbb{R}^d} f(\boldsymbol{x}), \tag{1}$$

其中,$\boldsymbol{x}$表示 d 维决策向量,$\mathcal{X}$表示决策空间,f表示目标函数.对应上述例子,$\boldsymbol{x}$可以表示成算法、传感器部署位置、实验配置、药物配方或食谱等,$f(\boldsymbol{x})$可表示为判断系统、环境、实验、配方、食谱等表现优劣的测度.

近年来,大数据应用的发展给物理学、生物学、环境生态学、计算机科学等领域以及军事、金融、通讯等行业带了巨大的生机.这些大数据应用通常都存在以下特点:大规模用户量、极其复杂的软件系统、大规模异构计算和分布式存储架构.这些复杂应用包含大量的设计决策,并且更为复杂,其优化目标不仅具有多峰、非凸、高维、决策空间巨大等常见特征,通常还具有黑箱和评估代价高昂等新特点.优化目标不存在明确的数学表达,并且需要花费高额代价才能观测到目标函数的返回值.例如:在研制某癌症的有效药物问题中,药物配方可以作为决策空间,药物效果[II]作为函数输出,临床试验作为评估药物效果的手段,目标是找到一个药物配方使得药物最大概率能够治愈病人.在该问题中,目标函数很难写成一个明确的数学表达式,评估函数过程可能会导致病人死亡.显然,这样的评估代价是巨大的.

针对具有以上特征的复杂设计问题,贝叶斯优化(Bayesian Optimization,简称 BO)是一个有效的解决方法[1].贝叶斯优化在不同的领域也称作序贯克里金优化(Sequential Kriging Optimization,简称 SKO)、基于模型的序贯优化(Sequential Model-Based Optimization,简称 SMBO)、高效全局优化(Efficient Global Optimization,简称 EGO).该方法是一种基于模型的序贯优化[III]方法,能够在很少的评估代价下得到一个近似最优解.贝叶斯优化已经应用于网页[2,3,4]、游戏[5]和材料设计[6]、推荐系统[7,8]、用户界面交互[9,10]、机器人步态[11]、导航[12]和嵌入式学习系统[13]、环境监控[14]、组合优化[15,16]、自动机器学习[17,18,19,20,21,22]、传感器网络[23,24]等领域,展示出令人瞩目的发展前景.

本文主要综述了贝叶斯优化方法研究和应用领域.组织结构如下:第 2 节引入贝叶斯优化的主要框架并深入分析其优化原理;第 3 节从模型选择角度介绍贝叶斯优化中两个核心组成部分:概率代理模型和采集函数;第 4 节介绍贝叶斯优化过程中涉及的近似和优化技术;第 5 节综述贝叶斯优化方法扩展及当前应用领域;第 6 节讨论其在未来发展中将面临的问题与挑战;第 7 节对其进行总结.

2 贝叶斯优化

概率模型已经成为当前人工智能、机器人学、机器学习等领域的主流方法[25].机器能够根据概率框架预测未来数据并且根据预测数据给出决策.这些问题的主要难点在于观测值具有不确定性,而概率模型能够对不确定性进行建模,有效地解决观测噪声问题.Ghahramani 指出贝叶斯优化是在概率机器学习和人工智能领域中几种最先进、最有希望的技术之一[25].

2.1 贝叶斯优化框架

贝叶斯优化是一个十分有效的全局优化算法,目标是找到(1)式中的全局最优解.贝叶斯优化有效地解决了序贯决策理论中经典的机器智能(machine-intelligence)问题:根据对未知目标函数 f 获取的信息,找到下一个评估位置,使得最快地达到最优解[26].例如:若已经评估得到三个不同输入 x_1、x_2、x_3 的目标函数值 y_1、y_2、y_3,则如何选择下一个评估点?贝叶斯优化框架能够在少数次评估下得到复杂目标函数的最优解,本质上因为贝叶斯优化框架使用代理模型拟合真实目标函数,并根据拟合结果主动选择最有“潜力”的评估点进行评估,避免不必要的采样,因此贝叶斯优化也称作主动优化(active optimization).同时,贝叶斯优化框架能够有效利用完整的历史信息来提高搜索效率.

I 本文只考虑最小化问题,最大化问题可简单通过取负号操作转换成最小化问题.

II 药物效果用药物能够治愈病人的概率大小来描述.

III 即在一次评估之后才进行下一次评估.

贝叶斯优化之所以称作“贝叶斯”,是因为优化过程中利用了著名的“贝叶斯定理”:

$$p(f \mid D_{1:t}) = \frac{p(D_{1:t} \mid f)p(f)}{p(D_{1:t})}, \tag{2}$$

其中,f 表示未知目标函数(或者表示参数模型中的参数), $D_{1:t} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), ..., (\boldsymbol{x}_t, y_t)\}$ 表示已观测集合, $\boldsymbol{x}_t$ 表示决策向量, $y_t = f(\boldsymbol{x}_t) + \varepsilon_t$ 表示观测值, ε_t 表示观测误差, $p(D_{1:t} \mid f)$ 表示 y 的似然分布,由于观测值存在误差,所以也称为“噪声”. $p(f)$ 表示 f 的先验概率分布,即对未知目标函数状态的假设. $p(D_{1:t})$ 表示边际化 f 的边际似然分布或者“证据”,由于该边际似然存在概率密度函数的乘积和积分,通常难得到明确的解析式.该边际似然在贝叶斯优化中主要用于优化超参数(hyper-parameter). $p(f \mid D_{1:t})$ 表示 f 的后验概率分布,后验概率分布描述通过已观测数据集对先验进行修正后未知目标函数的置信度.

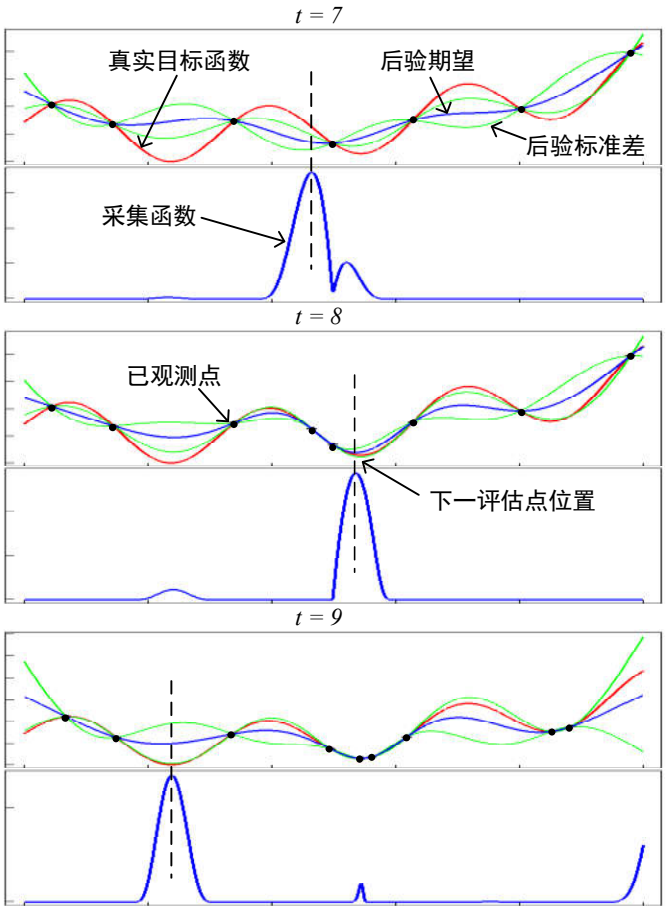

Fig.1 An example of using Bayesian Optimization on a 1D function $f(x) = (x-0.3)^2 + 0.2 \times \sin(20 \times x)$

图 1 贝叶斯优化在一维函数 $f(x) = (x-0.3)^2 + 0.2 \times \sin(20 \times x)$ 上的示例

贝叶斯优化框架主要包含两个核心部分:概率代理模型(probabilistic surrogate model)和采集函数(acquisition function).概率代理模型包含:先验概率模型和观测模型.先验概率模型即 $p(f)$.观测模型描述观测数据生成的机制,即似然分布 $p(D_{1:t} \mid f)$.更新概率代理模型意味着根据(2)式得到包含更多数据信息的后验概率分布 $p(f \mid D_{1:t})$.采集函数是根据后验概率分布构造的,通过最大化采集函数来选择下一个最有“潜力”的评估点.

同时,有效的采集函数能够保证选择的评估点序列使得总损失(loss)最小.损失有时表示为 regret:

$$r_t = |y^* - y_t|, \tag{3}$$

或累计 regret:

$$R_t = \sum_{i=1}^{t} r_i, \tag{4}$$

其中 y^* 表示当前最优解.

贝叶斯优化框架是一个迭代过程,主要包含三个步骤:第一步,根据最大化采集函数来选择下一个最有“潜力”的评估点 $\boldsymbol{x}_t$;第二步,根据选择的评估点 $\boldsymbol{x}_t$ 评估目标函数值 $y_t = f(\boldsymbol{x}_t) + \varepsilon_t$;第三步,把新得到的输入-观测值对 $\{\boldsymbol{x}_t, y_t\}$ 添加到历史观测集 $D_{1:t-1}$ 中,并且更新概率代理模型,为下一次迭代做准备.算法 1 为贝叶斯优化框架伪代码.

算法 1 贝叶斯优化框架

1: **for** $t = 1, 2, \ldots$ **do**
2: 最大化采集函数得到下一个评估点: $\boldsymbol{x}_t = \arg\max_{\boldsymbol{x}\in\mathcal{X}} \alpha(\boldsymbol{x}|D_{1:t-1})$;
3: 评估目标函数值: $y_t = f(\boldsymbol{x}_t) + \varepsilon_t$;
4: 整合数据: $D_t = D_{t-1} \cup \{\boldsymbol{x}_t, y_t\}$, 并且更新概率代理模型;
5: **end for**

图 1 为贝叶斯优化框架应用在一维函数 $f(x) = (x-0.3)^2 + 0.2 \times \sin(20 \times x)$ 上三次迭代的示例.横坐标表示决策空间,即 $\boldsymbol{x}$ 取值范围.红色曲线表示未知目标函数真实形状;上方蓝色曲线表示加入信息后的后验期望,即预测的函数曲线;两条绿色曲线表示两倍后验标准差,即不确定性.下方蓝色曲线表示采集函数,黑色虚线表示采集函数最大化的位置,即下一评估点位置.实心点表示已观测点.从上到下依次为第 7、8、9 次迭代,从图中可以看到,每次迭代选择的评估点都是上次迭代采集函数最大化位置.在迭代过程中,预测函数形状和采集函数形状都在变化,这是因为概率代理模型接收新数据后在不断更新.

2.2 贝叶斯优化原理

当优化危险化学试剂成分时,错误的试剂成分融合可能发生毁灭性的爆炸;当优化药物配方时,潜在致命的药物配方可能导致临床病人死亡;当优化航天飞机零部件配置时,不科学的零部件尺寸、结构配置可能导致航天飞机的运行不稳定甚至发生严重的航天事故.由于对这些优化目标进行评估时会花费大量的时间、费用甚至危害生命,因此在优化时通常希望在少量评估代价下得到满意解.相比其他无模型(model-free)优化算法(如进化计算和局部搜索等)关注于对求解效率提升,贝叶斯优化更侧重于减少评估代价,保证其能够仅经过少数次目标函数评估即可得到近优解.

在最优化采集函数的前提下,贝叶斯优化能够在理论上保证最终收敛[1].直观上,这是因为迭代过程中每次迭代都采样最有“潜力”的点进行评估,只要保证足量的迭代次数算法最终一定会收敛到全局最优解.当贝叶斯优化选择高斯过程代理黑箱函数并使用置信边界主动选择策略时,Srinivas 等人证明该方法能保证公式(4)对迭代次数 t 是亚线性的,即当迭代次数 t 趋于无穷时,公式(4)趋于 0[23].具体见 3.2.2 节.

Table 1 The features of optimization algorithms

表 1 几种优化算法特点对比

	最小代价	利用先验知识	弱假设	参数引入不确定性	主动选择策略
贝叶斯优化	√	√	√	√	√
K 摇臂赌博机	×	√	×	√	√
进化计算	×	×	√	×	×
局部搜索	×	×	√	×	×

表 1 为贝叶斯优化与其他优化算法特点比较.进化计算、局部搜索等无模型优化算法也能优化黑箱函数.这些方法通过选择、交叉、变异等操作模拟生物进化和群体智能过程,或者根据邻域信息探索最优解. 相比于它们,贝叶斯优化利用概率模型代理复杂黑箱函数.概率模型中引入待优化目标的先验知识,使模型更准确地满足黑箱函数的行为,有效地减少不必要的采样.K 摇臂赌博机等需要利用各决策之间相互独立等强假设. 与

之相比,贝叶斯优化在使用高斯过程代理黑箱函数时能够仅通过一致连续或利普希茨连续(Lipschitz continuity)等局部平滑性弱假设即可得到满意结果.局部平滑性等弱假设更符合实际问题,并且能够使贝叶斯优化有效利用局部邻近信息进行更准确的推断,从而更准确的选择“潜力”点.贝叶斯优化为参数引入不确定性,通过样本修正参数先验,并且在参数优化时利用贝叶斯方法,考虑参数的先验分布,通过平均得到参数,因此相比使用最大似然估计拟合参数等方法,该方法不易发生“过拟合”.贝叶斯优化通过主动选择策略来确定下一个最有“潜力”的评估点.相比无模型优化方法的随机跳转或邻域搜索策略,主动选择策略利用历史信息和不确定性,通过最大化根据模型后验分布构造的采集函数,能够有效地平衡宽度搜索(探索不确定性区域获取更多未知信息)与深度搜索(利用已有信息寻找当前最优)之间的关系,从而减少不必要的目标函数评估.

虽然贝叶斯优化具有多方面优势,但该方法仍存在以下局限性:1)无模型优化算法不需要考虑模型更新问题,而贝叶斯优化在更新概率代理模型时需要高昂的计算开销.如:使用高斯过程代理黑箱函数时,模型更新的时间复杂度为立方阶.一些研究采用近似技术和并行方法降低模型复杂度,提高计算效率,以缓解更新概率模型计算开销大的问题,具体见 4.1 节和 5.1.2 节.2)相比无模型的优化方法,贝叶斯优化需要谨慎地选择模型和先验.使用贝叶斯方法解决具体问题时,需要根据问题背景和专家知识选择合适的概率模型来代理黑箱函数.为贝叶斯优化选择合适的概率代理模型甚至比选择恰当的采集函数更为重要.目前,还不存在一种通用的方法为贝叶斯优化选择合适的代理模型和先验分布,都是采取具体问题具体分析的策略.

根据以上特点分析,贝叶斯优化适合求解优化目标存在多峰、非凸、黑箱、存在观测噪音并且评估代价高昂等特点的问题.例如:危险化学试剂实验、危害生命的药物测试、航空航天测试等等.但需要我们根据具体问题选择合适的模型代理模型和采集策略,才能充分发挥贝叶斯优化方法的潜力.

3 模型选择

贝叶斯优化框架有两个关键部分:1)使用概率模型代理原始评估代价高昂的复杂目标函数;2)利用代理模型的后验信息构造主动选择策略,即采集函数.在实际应用中,需要针对具体问题选择合适的模型.本节介绍贝叶斯优化中常用的概率代理模型和采集函数.在本节最后汇总常用概率代理模型和采集函数,并系统介绍各自方法的优劣及适用范围.

3.1 概率代理模型

概率代理模型用于代理未知目标函数,从假设先验开始,通过迭代地增加信息量、修正先验,从而得到更加准确的代理模型.概率代理模型根据模型的参数个数是否固定分为:参数模型和非参数模型.

3.1.1 参数模型

参数个数固定的概率模型称作参数模型.该模型在数据量增加和优化过程中,参数个数始终保持不变.使用 $\boldsymbol{w}$ 表示概率代理模型中的参数.由于 $\boldsymbol{w}$ 是模型中隐变量,假设其服从先验概率分布 $p(\boldsymbol{w})$.在参数 $\boldsymbol{w}$ 条件下观测数据的似然分布为 $p(D_{1:t}\mid \boldsymbol{w})$.根据“贝叶斯”定理,可以得到:

$$p(\boldsymbol{w}\mid D_{1:t})=\frac{p(D_{1:t}\mid \boldsymbol{w})p(\boldsymbol{w})}{p(D_{1:t})}, \tag{5}$$

$p(\boldsymbol{w}\mid D_{1:t})$ 表示经过观测数据 $D_{1:t}$ 修正后 $\boldsymbol{w}$ 的后验概率分布.注意到边际似然 $p(D_{1:t})$ 与 $\boldsymbol{w}$ 无关,通常用于优化超参数或为常量.由于(5)式分子是两个概率密度函数的乘积,所以后验概率分布一般难以得到封闭解.通常的解决方法是为参数 $\boldsymbol{w}$ 选择一个针对似然分布的共轭先验分布,使得到的后验概率分布与先验分布具有相同表达形式,便于计算.下面给出贝叶斯优化中几种常见的参数模型.

1) 贝塔-伯努利(Beta-Bernoulli)模型

首先讨论最简单的概率代理模型:贝塔-伯努利模型.再次用药物设计问题举例,假设存在 K 种药物,希望治愈患者,但是前提不知道哪种药物是有效的,并且药物有效性只能通过临床试验评估.目标是识别有效药物去治愈患者.这里首先假设所有药物的有效性相互独立,即不能通过观察一种药物的有效性推断出另一种药物的有效性.定义参数为 $\boldsymbol{w}$ 药物的有效概率.让患者服用一种药物后,假设能够观测到的 y 值只有两个状态,即

$y \in \{0,1\}$.也就是观测模型 $p(D_{1:t} \mid \boldsymbol{w})$ 为伯努利分布.为方便计算,假设 $\boldsymbol{w}$ 的先验分布 $p(\boldsymbol{w})$ 为贝塔分布[1]:

$$p(\boldsymbol{w} \mid \alpha, \beta) = \prod_{i=1}^{K} \mathrm{Beta}(w_i \mid \alpha, \beta), \tag{6}$$

其中, $\mathrm{Beta}(w \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} w^{\alpha-1}(1-w)^{\beta-1}$, α, β 为超参数, $\Gamma(.)$ 为伽玛函数.根据共轭性质,容易得到 $\boldsymbol{w}$ 的后验概率分布:

$$p(\boldsymbol{w} \mid D_{1:t}) = \prod_{i=1}^{K} \mathrm{Beta}(w_i \mid \alpha + n_{i,1}, \beta + n_{i,0}), \tag{7}$$

其中, $n_{i,1}$ 表示服用第 i 种药物成功治愈人数, $n_{i,0}$ 表示服用第 i 种药物导致死亡人数.

贝塔-伯努利模型不仅可以应用在药物设计问题,也可以应用在 A/B 测试[3]、推荐系统[4]等领域.

2) 线性(Linear)模型

在许多应用中,通常假设各决策之间相互独立.如在网页设计中,考虑页面布局、字体大小、颜色、按钮样式等 5 种因素,并且每种因素包含 5 种选择,因此总共有 625 种页面配置.若使用贝塔-伯努利模型,需要假设每种配置相互独立,因此为保证有效性每种配置都需要至少一次评估.因此,该方法不适合解决决策空间庞大的问题.然而,通过建立线性模型捕获各配置之间的关系,根据一种配置的表现来推断其他配置的表现,能够达到减少评估次数的目的[1].

在线性模型中,首先假设每种配置 i 都存在一个 d 维的特征向量 $\boldsymbol{m}_i \in \mathbb{R}^d$.在经过 t 次实验后,可以得到一个 $t \times d$ 的决策矩阵 $\boldsymbol{M}$, $\boldsymbol{M}$ 的第 j 行表示第 j 次实验所选配置对应的特征向量.定义目标函数为 $f: \mathbb{R}^d \to \mathbb{R}$,使任意配置都能通过 f 得到一个实数反馈, f 表达式如下:

$$f_{\boldsymbol{w}}(\boldsymbol{m}_i) = \boldsymbol{m}_i^T \boldsymbol{w}, \tag{8}$$

其中, $\boldsymbol{w}$ 表示权值向量.观测量 $y_i = f_{\boldsymbol{w}}(\boldsymbol{m}_i) + \varepsilon_i$,观测模型的形式取决于具体数据性质.假设噪声 ε_i 满足独立同分布: $p(\varepsilon_i) = \mathcal{N}(0, \sigma_i^2)$ [II].则 y_i 的似然分布为高斯分布:

$$p(y_i \mid \boldsymbol{w}, \sigma) = \mathcal{N}(\boldsymbol{m}_i^T \boldsymbol{w}, \sigma_i^2). \tag{9}$$

定义 $\boldsymbol{y}$ 为 t 维观测向量.由于高斯似然,方便起见假设参数 $\boldsymbol{w}, \sigma$ 服从 Normal-Inverse-Gamma 分布:

$$p(\boldsymbol{w}, \sigma \mid \boldsymbol{\mu}_0, \boldsymbol{V}_0, \alpha_0, \beta_0) = \mathcal{NIG}(\boldsymbol{\mu}_0, \boldsymbol{V}_0, \alpha_0, \beta_0) = \left|2\pi\sigma^2 \boldsymbol{V}_0\right|^{-\frac{1}{2}} \times \frac{{\beta_0}^{\alpha_0}}{\Gamma(\alpha_0)\sigma^{2\alpha_0+2}} \times \exp\left(-\frac{(\boldsymbol{w}-\boldsymbol{\mu}_0)^T \boldsymbol{V}_0^{-1}(\boldsymbol{w}-\boldsymbol{\mu}_0) + 2\beta_0}{2\sigma^2}\right), \tag{10}$$

其中, $\boldsymbol{\mu}_0, \boldsymbol{V}_0, \alpha_0, \beta_0$ 为超参数, $\boldsymbol{\mu}_0$ 为 d 维向量, $\boldsymbol{V}_0$ 为 $d \times d$ 的矩阵.由于 Normal-Inverse-Gamma 分布与高斯分布共轭,容易得到 $\boldsymbol{w}, \sigma$ 的后验分布:

$$p(\boldsymbol{w}, \sigma \mid D_{1:t}) = \mathcal{NIG}(\boldsymbol{\mu}_t, \boldsymbol{V}_t, \alpha_t, \beta_t), \tag{11}$$

其中,

$$\boldsymbol{\mu}_t = \boldsymbol{V}_t(\boldsymbol{V}_0^{-1}\boldsymbol{\mu}_0 + \boldsymbol{M}^T \boldsymbol{y}), \tag{12}$$

$$\boldsymbol{V}_t = (\boldsymbol{V}_0^{-1} + \boldsymbol{M}^T \boldsymbol{M})^{-1}, \tag{13}$$

$$\alpha_t = \alpha_0 + \tfrac{t}{2}, \tag{14}$$

$$\beta_t = \beta_0 + \tfrac{1}{2}(\boldsymbol{\mu}_0^T \boldsymbol{V}_0^{-1} \boldsymbol{\mu}_0 + \boldsymbol{y}^T \boldsymbol{y} - \boldsymbol{\mu}_t^T \boldsymbol{V}_t^{-1} \boldsymbol{\mu}_t). \tag{15}$$

公式(8)仅考虑线性关系,然而在大多数实际问题中, $\boldsymbol{m}_i$ 与 $f_{\boldsymbol{w}}(\boldsymbol{m}_i)$ 通常存在非线性关系.此时,可以通过使用 K 个非线性基函数 $\varphi_k: \mathbb{R}^d \to \mathbb{R}$ 来增加模型表示能力.即:

$$f_{\boldsymbol{w}}(\boldsymbol{m}_i) = (\varphi_1(\boldsymbol{m}_i), \varphi_2(\boldsymbol{m}_i), ..., \varphi_K(\boldsymbol{m}_i))\boldsymbol{w}, \tag{16}$$

其中,权重参数 $\boldsymbol{w}$ 为 K 维向量.比较常用的基函数有:径向基函数 $\varphi_k(\boldsymbol{m}_i) = \exp\left\{-\frac{1}{2}(\boldsymbol{m}_i - \boldsymbol{z}_k)^T \boldsymbol{\Lambda}(\boldsymbol{m}_i - \boldsymbol{z}_k)\right\}$ 和傅里叶基函数 $\varphi_k(\boldsymbol{m}_i) = \exp\left\{-a\boldsymbol{m}_i^T \boldsymbol{z}_k\right\}$ 等,其中 $\boldsymbol{z}_k$, $\boldsymbol{\Lambda}$ 和 a 是超参数.

3) 广义线性(Generalized Linear)模型

上面提到的线性模型能够捕获决策之间的关系,但是仅考虑实数型的观测量.为了推广线性模型处理其它

[I] 伯努利分布与贝塔分布是共轭的.

[II] $\mathcal{N}(.)$ 表示高斯分布概率密度函数.

类型的观测量(如:整型).Neider 等人提出广义线性模型(GLMs),通过 link function 把观测量从观测量空间映射到实数空间,使得能够处理的观测量类型更加灵活[27].

3.1.2 非参数模型

在机器学习中,高度灵活的模型通常能够得到满意的预测效果.这是因为这些模型具有高可扩展性等特点.一般有两种方法扩展模型的灵活性:(1)使参数模型拥有比数据集更多的参数.例如:目前用于翻译英语和法语的神经网络拥有 3 亿 8 千 4 百万个参数[28];(2)使用非参数模型.在非参数模型中,模型的参数随着数据量的增加而增加,甚至存在无限多个参数.因此,相比参数固定的参数模型,非参数模型更加灵活,并且使用贝叶斯方法不易发生“过拟合”[25].

1) 高斯过程

高斯过程(Gaussian Processes,简称 GPs)是常用的一种非参数模型.目前,高斯过程已经被广泛应用在回归、分类以及许多需要推断黑箱函数的领域中[29].GaussianFace[30]是高斯过程在人脸识别上的应用.该应用在人脸识别领域的表现胜过其他深度学习方法甚至人类.通常情况下,神经网络和高斯过程之间有这样一个联系:存在无限多个隐层单元的神经网络等价于高斯过程[31].

高斯过程是多元高斯概率分布的范化[29].一个高斯过程由一个均值函数 $m:\mathcal{X}\to\mathbb{R}$ 和一个半正定[I]的协方差函数 $k:\mathcal{X}\times\mathcal{X}\to\mathbb{R}$ 构成:

$$f(\boldsymbol{x})\sim\mathcal{GP}(m(\boldsymbol{x}),k(\boldsymbol{x},\boldsymbol{x}')), \tag{17}$$

其中,均值函数 $m(\boldsymbol{x})=\mathbb{E}[f(\boldsymbol{x})]$,协方差函数 $k(\boldsymbol{x},\boldsymbol{x}')=\mathbb{E}[(f(\boldsymbol{x})-m(\boldsymbol{x}))(f(\boldsymbol{x}')-m(\boldsymbol{x}'))]$.简单起见,通常设置均值函数 $m(\boldsymbol{x})=0$.

高斯过程是一个随机变量的集合,存在这样的性质:任意有限个随机变量都满足一个联合高斯分布[29].首先假设一个 0 均值的先验分布 $p(\boldsymbol{f}\mid\boldsymbol{X},\boldsymbol{\theta})$:

$$p(\boldsymbol{f}\mid\boldsymbol{X},\boldsymbol{\theta})=\mathcal{N}(0,\boldsymbol{\Sigma}), \tag{18}$$

其中,$\boldsymbol{X}$表示训练集 $\{\boldsymbol{x}_1,\boldsymbol{x}_2,...,\boldsymbol{x}_t\}$,$\boldsymbol{f}$表示未知函数$f$的函数值集合 $\{f(\boldsymbol{x}_1),f(\boldsymbol{x}_2),...,f(\boldsymbol{x}_t)\}$,$\boldsymbol{\Sigma}$ 表示 $k(\boldsymbol{x},\boldsymbol{x}')$ 构成的协方差矩阵[II],$\boldsymbol{\theta}$ 表示超参数.

当存在观测噪声时,即: $y=f(\boldsymbol{x})+\varepsilon$,且假设噪声 ε 满足独立同分布的高斯分布: $p(\varepsilon)=\mathcal{N}(0,\sigma^2)$.从而得到似然分布:

$$p(\boldsymbol{y}\mid\boldsymbol{f})=\mathcal{N}(\boldsymbol{f},\sigma^2\boldsymbol{I}), \tag{19}$$

其中,$\boldsymbol{y}$ 表示观测值集合 $\{y_1,y_2,...,y_t\}$.

根据公式(18)和(19)可以得到边际似然分布:

$$p(\boldsymbol{y}\mid\boldsymbol{X},\boldsymbol{\theta})=\int p(\boldsymbol{y}\mid\boldsymbol{f})p(\boldsymbol{f}\mid\boldsymbol{X},\boldsymbol{\theta})d\boldsymbol{f}=\mathcal{N}(\boldsymbol{0},\boldsymbol{\Sigma}+\sigma^2\boldsymbol{I}). \tag{20}$$

通常通过最大化该边际似然分布优化超参数 $\boldsymbol{\theta}$.

根据高斯过程的性质,存在如下联合分布:

$$\begin{bmatrix}\boldsymbol{y}\\ \boldsymbol{f}_*\end{bmatrix}\sim N(\boldsymbol{0},\begin{bmatrix}\boldsymbol{\Sigma}+\sigma^2\boldsymbol{I} & K_*\\ K_*^T & K_{**}\end{bmatrix}), \tag{21}$$

其中, $\boldsymbol{f}_*$ 表示预测函数值, $\boldsymbol{X}_*$ 表示预测输入, $K_*^T=\{k(\boldsymbol{x}_1,\boldsymbol{X}_*),k(\boldsymbol{x}_2,\boldsymbol{X}_*),...,k(\boldsymbol{x}_t,\boldsymbol{X}_*)\}$, $K_{**}=k(\boldsymbol{X}_*,\boldsymbol{X}_*)$.

根据公式(21)可以容易地得到预测分布:

$$p(\boldsymbol{f}_*\mid\boldsymbol{X},\boldsymbol{y},\boldsymbol{X}_*)=\mathcal{N}(<\boldsymbol{f}_*>,\mathrm{cov}(\boldsymbol{f}_*)), \tag{22}$$

$$<\boldsymbol{f}_*>=K_*^T[\boldsymbol{\Sigma}+\sigma^2\boldsymbol{I}]^{-1}\boldsymbol{y}, \tag{23}$$

$$\mathrm{cov}(\boldsymbol{f}_*)=K_{**}-K_*^T[\boldsymbol{\Sigma}+\sigma^2\boldsymbol{I}]^{-1}K_*, \tag{24}$$

I 一个有效的协方差函数必须是半正定的[29].

II $\Sigma_{i,j}=k(\boldsymbol{x}_i,\boldsymbol{x}_j)$.

其中, $< \boldsymbol{f}_* >$ 表示预测均值, $\mathrm{cov}(\boldsymbol{f}_*)$ 表示预测协方差[I].

先验均值函数表示目标函数期望的偏移量.为了增加模型的解释性同时先验信息方便表达时,可以明确地指定先验均值函数 $m(\boldsymbol{x})$.这时预测协方差与公式(24)相同,预测均值变为:

$$< \boldsymbol{f}_* >= m(\boldsymbol{X}_*) + K_*^T[\boldsymbol{\Sigma} + \sigma^2 \boldsymbol{I}]^{-1}(\boldsymbol{y} - m(\boldsymbol{X}_*)). \tag{25}$$

然而在实际应用中,指定一个明确的、合理的先验均值函数十分困难[29].所以为了简便,通常假设先验均值函数为恒 0 函数,即: $m(\boldsymbol{x}) = 0$.注意到,当 $m(\boldsymbol{x}) = 0$ 时,通过数据修正后的后验均值(23)并不限制为 0,因此该假设对后验准确性几乎不影响.

在高斯过程中存在这样一致连续或利普希茨连续的平滑性假设:当输入点 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 之间距离特别近时,相应观测值 y_i 和 y_j 是相似的.因此,预测样本附近的训练样本能够提供更多信息.协方差函数是高斯过程中计算两个数据点之间相似性的函数,它指定了未知目标函数的平滑性和振幅.因此,对协方差函数的选择直接影响着高斯过程与数据性质之间的匹配程度.

在实际应用中,只有选择合适的协方差函数才能保证得到理想的预测效果.协方差函数一般分为平稳(stationary)[II]协方差函数和非平稳协方差函数.若目标函数具有非平稳性,可以直接使用非平稳协方差函数[32]或者通过把目标函数分离成多个平稳区域,并在每个区域内使用平稳协方差函数[33]的方法处理.

常用的平稳协方差函数有:平方指数(Squared Exponential)协方差函数、指数(Exponential)协方差函数和 Matérn 协方差函数等等.

Matérn 协方差函数簇是一类高灵活性的协方差函数.具体函数表达式如下:

$$k_{\text{Matérn-}\nu}(r) = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}r}{l}\right)^{\nu} K_{\nu}\left(\frac{\sqrt{2\nu}r}{l}\right) \quad , \tag{26}$$

其中, $r = |\boldsymbol{x} - \boldsymbol{x}'|$, ν 为平滑参数, l 为尺度参数, K_ν 为第二类变形贝塞尔函数[34].

从使用 Matérn 协方差函数的高斯过程中采样的目标函数 $f(\boldsymbol{x})$ 是 $\lfloor \nu - 1 \rfloor$ 次均方可微的.在机器学习领域比较感兴趣是当 ν=1/2、ν=3/2、ν=5/2、 $\nu \to \infty$ 时的情况.表 2 列出几种常用的 Matérn 协方差函数.

Table 2 Common Matérn covariance functions

表 2 常用 Matérn 协方差函数

ν	具体表达式
1/2	$k_{\text{Matérn-}\frac{1}{2}}(r) = \exp(-\frac{r}{l})$
3/2	$k_{\text{Matérn-}\frac{3}{2}}(r) = (1 + \frac{\sqrt{3}r}{l})\exp(-\frac{\sqrt{3}r}{l})$
5/2	$k_{\text{Matérn-}\frac{5}{2}}(r) = (1 + \frac{\sqrt{5}r}{l} + \frac{5r^2}{3l^2})\exp(-\frac{\sqrt{5}r}{l})$
$\nu \to \infty$	$k_{\text{Matérn-}\infty}(r) = \exp(-\frac{r^2}{2l^2})$

当 ν=1/2 时,Matérn 协方差函数也称作指数协方差函数,对应的过程均方连续但不满足均方可微,所以局部存在较大波动.当 $\nu \to \infty$ 时,对应的协方差函数也称作平方指数协方差函数或者高斯协方差函数,对应的过程无限均方可微,高度平滑.

当 $\boldsymbol{x}$ 为 d 维时,可以为每个维度指定一个尺度参数 l_j,其中 $1 \le j \le d$.这样的协方差函数实现了自动相关确定(Automatic Relevance Determination,简称 ARD)[31].因为尺度参数的倒数决定了该维度的相关性,即:若 l_j 很大,则该维度几乎独立,这样能够有效地识别并去除不相关维度.

当 $\boldsymbol{x}$ 属于离散类型或分类类型时,使用汉明(Hamming)协方差函数的高斯过程通常具有良好的表现.并且当 $\boldsymbol{x}$ 为二元维度时,使用带自动相关确定的高斯协方差函数能够有效地计算二元维度之间的加权汉明距离.

也可根据问题特性使用有效协方差组合.更多的协方差函数在[29]中详细介绍.

I 若 f_* 为标量,则 $< f_* >$ 表示均值, $\mathrm{cov}(f_*)$ 表示协方差;若 $\boldsymbol{f}_*$ 为向量,则 $< \boldsymbol{f}_* >$ 表示均值向量, $\mathrm{cov}(\boldsymbol{f}_*)$ 表示协方差矩阵.

II 平稳的协方差函数满足: $k(\boldsymbol{x}, \boldsymbol{x}') = k(\boldsymbol{x} - \boldsymbol{x}')$.

2) 随机森林

随机森林回归是一种十分适合并行化的回归方法[35].该方法属于集成学习,即通过组合多个弱学习器来提高预测精度.随机森林回归构造多棵决策树,每棵决策树通过从训练数据中有放回的采样进行训练.当需要预测时,把采样点输入到每棵决策树中,并得到每棵树的预测均值,然后通过投票机制得到最终预测结果.

与高斯过程高昂更新代价相比,随机森林方法具有极其优秀的计算效率.由于其计算高效性和对大规模数据集的有效性,该方法已成功应用于自动算法配置领域[33].

虽然随机森林回归在训练数据附近能够快速得到高精度预测,但是在远离训练数据时的预测效果通常很差,并且该方法的响应面是非连续、不可微的,因此不能对其使用基于梯度的优化方法.

3) 深度神经网络

深度神经网络通常指层数超过 2 层的神经网络.虽然具有无限多个隐层单元的神经网络等价于高斯过程,但该神经网络具有无穷多个参数,无法训练.为了减少参数个数,一种常用的方法就是增加神经网络的深度.

近年来,由于其优越的性能,深度神经网络已成功应用于语音识别[36]、机器视觉[37]等领域.在贝叶斯优化领域中,深度神经网络同样得到重视.一些研究者通过使用深度神经网络代理未知目标函数,来提升模型处理大规模数据的能力[38,39].然而,若想得到理想效果的函数近似,需要合理地设计神经网络架构,如:层数、每层的神经元个数等.如何设计合理的神经网络架构仍是具有挑战性的问题.

3.2 采集函数

前一节介绍了代理复杂黑箱目标函数的概率模型,并介绍了如何结合新样本进行模型更新.本节介绍在贝叶斯优化中选择下一个评估点的主动策略:采集函数.所谓采集函数就是从输入空间 $\mathcal{X}$、观测空间 $\mathbb{R}$ 和超参数空间 Θ 映射到实数空间的函数 $\alpha:\mathcal{X}\times\mathbb{R}\times\Theta\to\mathbb{R}$.该函数由已观测数据集 $D_{1:t}$ 得到的后验分布构造,并通过对其最大化指导选择下一个评估点 $\boldsymbol{x}_{t+1}$:

$$\boldsymbol{x}_{t+1}=\max_{\boldsymbol{x}\in\mathcal{X}}\alpha_t(\boldsymbol{x};D_{1:t}). \tag{27}$$

为方便描述,本节不考虑采集函数对超参数的依赖,对超参数的优化将在 4.2.1 节介绍.图 2 为几种常用采集函数对比.

3.2.1 基于提升的策略

基于提升的策略偏好选择对于当前最优目标函数值有所提升[I]的位置作为评估点.

PI(Probability of Improvement)量化了 $\boldsymbol{x}$ 的观测值可能提升当前最优目标函数值的概率[40].PI 的采集函数为:

$$\alpha_t(\boldsymbol{x};D_{1:t})=p(f(\boldsymbol{x})\le v^*-\xi)=\phi\left(\frac{v^*-\xi-\mu_t(\boldsymbol{x})}{\sigma_t(\boldsymbol{x})}\right), \tag{28}$$

其中,v^*表示当前最优函数值,$\phi(.)$为标准正态分布累积密度函数,ξ 为平衡参数[II].如图 2 所示,在当前最优解附近时,公式(28)的取值很大,并且在远离当前最优解时取值很小.对参数 ξ 的调整能够一定程度上解决陷入局部最优的问题.当 ξ 较大时, $f(\boldsymbol{x})\le v^*-\xi$ 在决策空间上的概率都较小,公式(28)整体平缓,此时 PI 策略更针对全局检索.反之,当 ξ 较小时,公式(28)整体相对挺拔,PI 策略更针对局部检索[41].

虽然 PI 策略能够选择提升概率最大的评估点,但是 PI 策略把所有提升看成是等量的,只反映了提升的概率而没有反映提升量的大小.

Močkus 等人提出一种新的基于提升的策略:EI(Expected Improvement)[42].EI 策略的采集函数为:

$$\alpha_t(\boldsymbol{x};D_{1:t})=\begin{cases}(v^*-\mu_t(\boldsymbol{x}))\phi\left(\frac{v^*-\mu_t(\boldsymbol{x})}{\sigma_t(\boldsymbol{x})}\right)+\sigma_t(\boldsymbol{x})\phi\left(\frac{v^*-\mu_t(\boldsymbol{x})}{\sigma_t(\boldsymbol{x})}\right), & \sigma_t(\boldsymbol{x})>0\\ 0, & \sigma_t(\boldsymbol{x})=0\end{cases} \tag{29}$$

I 这里的提升指比当前目标函数值小.

II 用于平衡局部和全局搜索之间关系的参数,一般人为设定.

其中, ϕ(.) 为标准正态分布概率密度函数.如图 2 所示,EI 策略选择的 $\boldsymbol{x}$ 与 PI 策略有所不同,因为 EI 策略的采集函数包含公式(28),既整合了提升概率又体现了不同的提升量.当然,EI 策略同样可以包含平衡参数 ξ 进一步处理局部和全局之间的关系[43]:

$$\alpha_t(\boldsymbol{x};D_{1:t})=\begin{cases}(v^*-\xi-\mu_t(\boldsymbol{x}))\phi(\frac{v^*-\xi-\mu_t(\boldsymbol{x})}{\sigma_t(\boldsymbol{x})})+\sigma_t(\boldsymbol{x})\phi(\frac{v^*-\xi-\mu_t(\boldsymbol{x})}{\sigma_t(\boldsymbol{x})}), & \sigma_t(\boldsymbol{x})>0\\ 0, & \sigma_t(\boldsymbol{x})=0\end{cases} \tag{30}$$

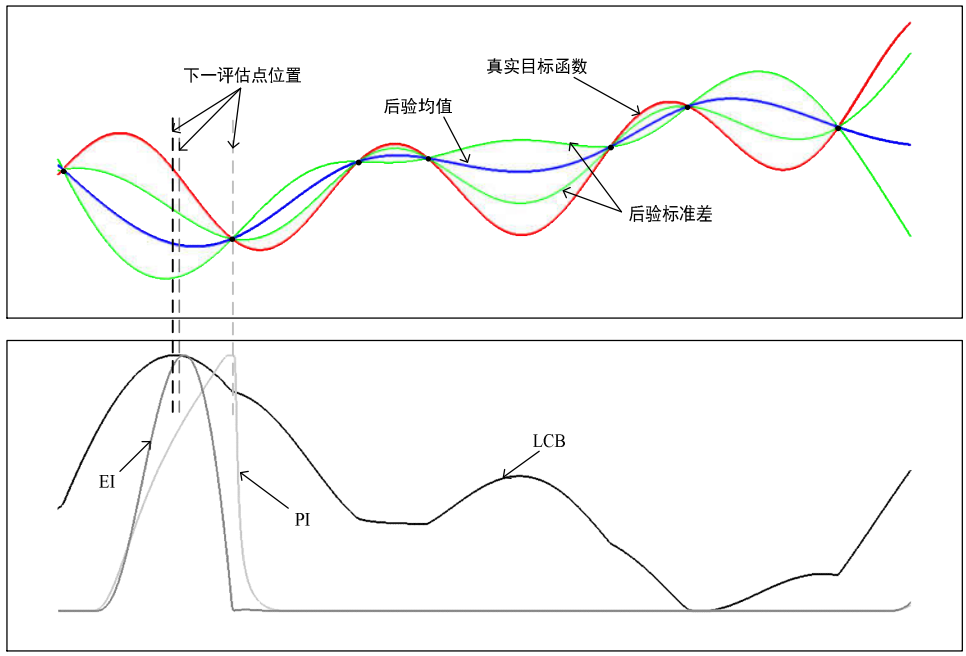

Fig.2 Comparison of several common acquisition functions

图 2 几种常用采集函数对比

3.2.2 置信边界策略

置信边界策略已经在 K 摇臂赌博机领域中广泛应用[44].Srinivas 等人在 2010 年提出一种针对高斯过程的置信边界策略:GP-UCB[23].UCB 表示置信上界,在求解目标函数最大值时,UCB 策略的采集函数为:

$$\alpha_t(\boldsymbol{x};D_{1:t})=\mu_t(\boldsymbol{x})+\sqrt{\beta_t}\sigma_t(\boldsymbol{x}). \tag{31}$$

然而当求解目标函数的最小值时,使用置信下界策略 LCB:

$$\alpha_t(\boldsymbol{x};D_{1:t})=-(\mu_t(\boldsymbol{x})-\sqrt{\beta_t}\sigma_t(\boldsymbol{x})), \tag{32}$$

其中,参数 β_t 平衡了期望和方差.

Srinivas 等人给出了相对于不同协方差函数参数 β_t 的具体表达式,同时理论证明了在 β_t 取这些值时公式(4)对迭代次数 t 是亚线性的[23].从公式(31)和公式(32)可以看出,当不确定性较大时,采集函数取值同样很大,如图 2 所示,LCB 采集函数在不确定性大的地方存在波峰,并且最大化 LCB 采集函数的 $\boldsymbol{x}$ 偏向置信下边界的最小值,这个偏向程度取决于参数 β_t 的大小.

3.2.3 基于信息的策略

首先定义在未知目标函数全局最优解 $\boldsymbol{x}^*$上的后验分布为: $p^*(\boldsymbol{x}\mid D_{1:t})$.

汤普森采样(Thompson Sampling,简称 TS)[45]是一个随机策略:从后验分布 $p^*(\boldsymbol{x}\mid D_{1:t})$ 中采样一个收益(reward)函数,然后选择具有最大收益的 $\boldsymbol{x}$ 作为下一个评估点.

考虑药物设计问题,参数 $\boldsymbol{w}$ 为药物有效概率.汤普森采样的采集函数为:

$$\alpha_t(x;D_{1:t})=\boldsymbol{w}_x\,, \tag{33}$$

其中,x 表示第 x 种药物,$\boldsymbol{w}_x$ 表示第 x 种药物的有效概率,$\boldsymbol{w}\sim p(\boldsymbol{w}\mid D_{1:t})$.因此下一个评估点选取使公式(33)最大化的第 x 种药物进行评估.直观来讲就是根据已观测数据集选择有效概率最大的药物进行临床试验.这样相比较每种药物做同样多的临床试验再统计有效概率而言,若某种药物具有高致死率,统计方法会导致大量不必要的死亡,而汤普森采样能够避免这一情况的发生,明显地减少临床试验的代价.

汤普森采样策略存在如下几个优势:1)没有多余的参数并且容易实现;2)由于根据后验分布随机采样选取最优,该方法很自然地平衡了寻找当前最优与探索新区域之间的关系;3)特别适合批量或者延迟的反馈情形[1].但是,汤普森采样由于其随机性导致具有强宽度搜索性质,因此在处理高维度问题时经常导致很差的性能[46].

前面提到的汤普森采样针对搜索空间是离散的,当搜索空间 $\mathcal{X}$ 为连续时,汤普森采样需要通过频域抽样技术[47]从高斯过程中近似采样收益函数.

熵搜索策略(Entropy Search,简称 ES)[46]主要思想是减少全局最优解 $\boldsymbol{x}^*$的不确定性,不同于汤普森采样,不是从 $p^*(\boldsymbol{x}\mid D_{1:t})$ 中采样,而是针对其不确定性进行检索.换句话说,ES 策略主要偏好选取能够对最优解$\boldsymbol{x}^*$提供最多信息的 $\boldsymbol{x}$.ES 策略的采集函数为:

$$\alpha_t(\boldsymbol{x};D_{1:t})=H(p(\boldsymbol{x}^*\mid D_{1:t}))-\mathbb{E}_{p(y\mid\boldsymbol{x},D_{1:t})}[H(p(\boldsymbol{x}^*\mid D_{1:t}\cup\{\boldsymbol{x},y\}))]\,, \tag{34}$$

其中,$H(p(x))=\int p(x)\log p(x)dx$ 表示 $p(x)$的微分熵,$p(y\mid\boldsymbol{x},D_{1:t})=\mathcal{N}(\mu_t(\boldsymbol{x}),\sigma_t^2(\boldsymbol{x})+\sigma^2)$.

公式(34)在实际中是不能精确计算的,因为计算 $p(\boldsymbol{x}^*\mid D_{1:t}\cup\{\boldsymbol{x},y\})$ 需要取大量不同的 $\boldsymbol{x}$,y 对,并且微分熵本身难以计算.因此计算公式(34)需要使用近似技术,如:蒙特卡洛采样技术[48].

Hernándezlobato 等人在 2014 年提出熵预测搜索(Predictive Entropy Search,简称 PES)[49].该方法利用 $\boldsymbol{x}^*$与 y 之间互信息(mutual information)的对称性重写公式(34)为:

$$\alpha_t(\boldsymbol{x};D_{1:t})=H(p(y\mid D_{1:t}))-\mathbb{E}_{p(\boldsymbol{x}^*\mid D_{1:t})}[H(p(y\mid D_{1:t},\boldsymbol{x},\boldsymbol{x}^*))]\,. \tag{35}$$

相比公式(34)依赖 $p(\boldsymbol{x}^*\mid D_{1:t}\cup\{\boldsymbol{x},y\})$,公式(35)依赖预测分布的熵,由于预测分布容易近似甚至精确计算,因此 PES 策略更方便计算.Hernándezlobato 等人通过实验验证了 PES 策略是目前熵搜索策略中最先进的技术之一[49].

3.2.4 组合策略

采用单一采集函数的贝叶斯优化算法不可能在所有问题上都表现出最好的性能[50].因此为了得到一个具有强鲁棒性的方法,Brochu 等人提出一种使用对冲策略组合多种采集函数(如:PI、EI、UCB 等等)的贝叶斯优化算法:GP-Hedge[50].GP-Hedge 算法在每次迭代中把从每种采集函数得到的候选点组成候选点集合,然后根据对冲策略从候选点集合中选取评估点.所谓对冲策略就是根据每个采集函数 i 的有效概率 $p(i)$选取评估点,概率 $p(i)$根据采集函数 i 的累计收益计算得到.该方法本质上是通过采集函数的历史效果来预测将来的效果.GP-Hedge 算法相比较使用单一采集函数的贝叶斯优化算法能够表现出更强的鲁棒性.

Shahriari 等人提出一种基于信息的组合策略 ESP(Entropy Search Portfolio)[46].ESP 策略与 GP-Hedge 不同之处在于该方法通过基于熵搜索方法选择能为全局最优解提供最多信息的候选点为评估点.ESP 策略同样具有高鲁棒性,并且相比对冲策略该方法能够容忍组合中存在表现差的采集函数.这是因为对冲策略初期可能做出错误选择,从而影响整体优化效率,而 ESP 策略克服了这一缺点.

3.3 常用模型汇总

利用贝叶斯优化解决实际问题时,选择合适的概率代理模型和采集函数是十分重要的.然而,在贝叶斯优化领域中没有统一而明确的选择标准,这仍是具有挑战性的开放问题.为了从总体上更加清晰地认识常用的代理模型和采集函数,我们总结了各自方法的优势、劣势以及其代表性应用和文献. 希望为相关研究者提供有益的参考，帮助他们在采用贝叶斯优化解决实际问题时选择合适的模型. 表 3 汇总了常用的概率代理模型以及各自的优势、劣势、适用范围和代表性应用. 表 4 汇总了常用的采集函数以及各自的优势、劣势和代表性文

献.

Table 3 A summary of common probabilistic surrogate models

表 3 常用概率代理函数汇总

类别	概率代理模型	优势	劣势	代表性应用
参数模型	贝塔-伯努利模型	简单,适用于观测量为二值的问题	需假设各决策之间独立,不适用决策空间庞大的问题	A/B 测试[3]、新闻推荐[4]
	线性模型	考虑各决策之间依赖关系,适用于决策空间庞大的问题	需预先为决策定义 d 维特征,仅能处理实数型观测量	传感器网络和自动算法配[22]
	广义线性模型	具有线性模型的优点,且能处理任意类型(如:整型)观测量	需预先为决策定义 d 维特征	A/B 测试[3]、新闻推荐[4]
非参数模型	高斯过程	高灵活性和可扩展性,理论上能代理任意线性/非线性函数	训练复杂度高,不适用具有大量已观测样本的问题,且需预先仔细选择领域相关的协方差函数	机器人控制[11,12],传感器网络[14,23],偏好学习[9],自动算法配置[17,18],自然语言处理[61,82]
	随机森林	计算效率高,适用于大规模数据集	预测精度与训练集高度依赖,且不能对其使用基于梯度的优化方法	自动算法配置[33]
	深度神经网络	可处理大规模数据	需预先设计合适的神经网络结构(如:每层的神经元个数,层数等)	自动算法配置[38,39]

Table 4 A summary of common acquisition functions

表 4 常用采集函数汇总

类别	采集函数	优势	劣势	代表性文献
基于提升的策略	PI	简单易推导	把提升看作等量,仅反映提升的概率而没有反映提升量的大小	[40,41]
	EI	参数少,既整合提升的概率又体现不同的提升量,并平衡了深度和宽度之间的关系	当使用基于梯度的方法优化时,需推导导数信息	[42,43]
置信边界策略	置信边界策略	简单,平衡了深度和宽度之间的关系	对参数 β_t 敏感	[23]
基于信息的策略	汤普森采样	无多余参数, 其随机性避免局部最优,适用于批量或延迟反馈的情形	具有强宽度搜索性质,不适用处理高维度问题	[45]
	熵搜索策略	利用熵定义精确减少最优解的不确定性	计算量高,且需使用近似技术	[46]
	熵预测策略	具有熵搜索策略优点,且方便计算	引入熵和期望的计算量	[49]
组合策略	GP-Hedge	强鲁棒性	每次迭代需计算所有采集函数,增加计算量,且优化初期可能做出错误选择	[50]
	ESP	强鲁棒性,对表现差的采集函数高容忍,并能克服初期错误选择	每次迭代需计算所有采集函数,同时引入熵的计算	[46]

4 近似与优化技术

上节介绍了多种常用的概率代理模型和采集函数.选择合适的模型之后,需要考虑如何对概率模型进行更新推断,并且如何优化超参数和采集函数.本节介绍贝叶斯优化过程中涉及的近似和优化技术.

4.1 近似技术

在贝叶斯优化中,概率代理模型更新是迭代优化过程中的核心步骤.概率代理模型更新是指根据整合的新样本推断出模型后验.

当模型先验与似然分布非共轭时,难以得到后验分布的封闭解.该情况可以采用变分贝叶斯(Variational Bayesian, VB)近似推断[51]或蒙特卡洛近似方法得到近似后验分布.并且这两种近似方法的好处在于能够处理任意类型的模型先验与似然,十分灵活.

当概率代理模型为高斯过程时,由于在推断后验分布时需要计算 $t\times t$ 协方差矩阵的逆,因此高斯过程的精

确推断需要的时间复杂度为立方阶.高斯过程精确推断不适用于当训练样本数 t 特别大的情况.在实际应用中,计算协方差矩阵的逆通常使用Cholesky分解[29]方法.该方法十分稳定,并且在超参数未更新时,只需计算并存储一次计算复杂度为 $O(t^3)$的 H 矩阵,而预测的时间复杂度仅为 $O(t^2)$.但是超参数在每次迭代都有可能更新,进而需要重新计算 H 矩阵.因此,在处理大数据集时,需要通过使用近似技术来权衡精度和计算复杂度之间的关系.

Table 5 Common approximation techniques

表 5 常用近似技术

方法	时间复杂度
精确 GP	$O(t^3)$
Cholesky 分解[29]	$O(t^2)$
SPGP[52,53]	$O(tm^2+m^3)$
SSGP[47]	$O(tm^2+m^3)$

表 5 列出几种针对高斯过程的常用近似技术和对应时间复杂度.SPGP(Sparse Gaussian Process Using Pseudo-inputs)方法属于降秩近似方法,引入 $m<t$ 个伪输入使得协方差矩阵的秩近似降为 m,从而明显减少求解协方差矩阵逆的计算量[52,53].[52]与[53]中的算法虽然时间复杂度相同,但是[52]方法中 m 个伪输入的位置是固定的,而[53]方法中的 m 个伪输入位置是改变的,即把 m 个伪输入的位置也当成参数,在优化超参数时一同进行优化,使得伪输入的位置能够根据当前数据集进行学习,增加模型弹性.但是当数据集较小时,由于参数相对较多,容易产生“过拟合”现象.SSGP(Sparse Spectrum Gaussian Process)方法的关键思想是稀疏化高斯过程的谱表示[47].博赫纳定理指出任何平稳的协方差函数都能表示成有限测度的傅里叶变换:

$$k(\boldsymbol{x},\boldsymbol{x}')=k(\boldsymbol{r})=\int \exp(2\pi i\boldsymbol{s}^T(\boldsymbol{x}-\boldsymbol{x}'))S(\boldsymbol{s})d\boldsymbol{s}\,. \tag{36}$$

之后通过蒙特卡洛方法采样 m 个样本频率(spectral points)近似公式(36)的积分从而得到近似的协方差函数.当数据集较小时,SSGP 同样易产生“过拟合”现象.更多的高斯过程近似方法参考[29]中第 8 章.

4.2 优化技术

4.2.1 超参数优化

概率代理模型中超参数的取值直接影响模型的预测效果.对这些超参数的优化是必要的.极端情况下,每次迭代都根据当前的数据重新学习所有超参数.这样的方法虽然能保证模型的准确率,但是对超参数的学习需要高昂的计算量,效率低下.当前常用的贝叶斯优化实现(如:BayesOPT[54])通常采用的方法是多次迭代之后(如 20 次迭代)重新学习超参数.

贝叶斯优化中一般对超参数处理有两种方法:点估计和近似边际化估计.

点估计方法有:1)通过第 II 类极大似然估计(type II maximum likelihood,简称 ML)对边际似然(20)最大化得到 $\hat{\boldsymbol{\theta}}_t^{ML}$;2)为超参数赋予先验 $p(\boldsymbol{\theta})$,然后通过“贝叶斯定理”得到:

$$p(\boldsymbol{\theta}\mid D_{1:t})=\frac{p(D_{1:t}\mid\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(D_{1:t})}\,. \tag{37}$$

最后通过最大后验估计(maximum a posteriori,简称 MAP)最大化公式(37)得到 $\hat{\boldsymbol{\theta}}_t^{MAP}$;3)最大化留一法交叉验证得到的似然均值(称作 leave-one-out likelihood 或 pseudo-likelihood),得到超参数估计 $\hat{\boldsymbol{\theta}}_t^{LOOCV}$.该方法对“过拟合”具有鲁棒性,通常能够提升模型质量.

根据上面估计出的 $\hat{\boldsymbol{\theta}}$ 能够得到具体的采集函数:

$$\hat{\alpha}_t(\boldsymbol{x})=\alpha(\boldsymbol{x};\boldsymbol{\theta}_t)\,. \tag{38}$$

在贝叶斯优化中,不确定性在指导宽度搜索时起到重要作用.然而,上面提到的点估计方法本质上不能捕获这些不确定性.因此,为了处理这一问题,通常对 $\boldsymbol{\theta}$ 进行边际化处理:

$$\alpha_t(\boldsymbol{x})=\int p(\boldsymbol{\theta}\mid D_{1:t})\alpha(\boldsymbol{x};\boldsymbol{\theta})d\boldsymbol{\theta}\,. \tag{39}$$

常使用蒙特卡洛方法近似得到公式(39)中的积分.即首先从后验分布 $p(\boldsymbol{\theta}\mid D_{1:t})$ 中采样 M 个 $\boldsymbol{\theta}$ 的样本,然后通过公式(40)得到近似积分:

$$\alpha_t(\boldsymbol{x}) \approx \sum_{i=1}^{M} \alpha(\boldsymbol{x};\boldsymbol{\theta}_t^{(i)}). \tag{40}$$

Osborne 等人[55]使用贝叶斯蒙特卡洛技术[56]近似公式(39)中的积分.该方法中 $\boldsymbol{\theta}$ 的样本不是从后验分布 $p(\boldsymbol{\theta}|D_{1:t})$ 中选取的,而是选取使得预测分布均值(23)和方差(24)不同的 M 个样本,然后对这 M 个样本加权求和近似:

$$\alpha_t(\boldsymbol{x}) \approx \sum_{i=1}^{M} \rho^{(i)} \alpha(\boldsymbol{x};\boldsymbol{\theta}_t^{(i)}), \tag{41}$$

其中,权值 $\rho^{(i)}$ 在[57]中详细描述.

4.2.2 采集函数优化

在 3.2 节介绍了几种常用采集函数,注意到在贝叶斯优化中通过这些采集函数选取下一个评估点时,需要通过最大化公式(27).然而采集函数通常是非凸、多峰的甚至在本质上比目标函数更难优化.但是相比较目标函数,采集函数的评估代价通常很小.因此为了不影响求解效率,通常需要尽量简单地设计采集函数.优化采集函数的方法称作辅助优化器.目前在贝叶斯优化研究中常用的辅助优化器有:离散化方法[18]、多启动拟牛顿爬山方法[18]、网格搜索[58]、无利普希茨常量的利普希茨优化(DIRECT)[59]、自适应协方差矩阵进化策略(CMA-ES)[60]和多启动局部搜索[31].这些辅助优化器通常存在如下缺点:1)很难判定这些辅助优化器是否找到采集函数的全局最优解,只有当选择的下一个评估点是采集函数的全局最优时,才能在理论上保证贝叶斯优化的收敛性[1];2)在两次连续的贝叶斯优化迭代中,采集函数可能没有显著的变化,因此每次都从头优化采集函数会浪费大量的计算资源.Wang 等人提出一种不需要精确地优化采集函数即可理论保证贝叶斯优化收敛的方法[61].该方法将乐观[1]优化算法与概率代理模型进行结合,忽略计算复杂黑箱函数的后验分布过程,通过扩展函数值大或者为上界的叶子节点构造一个空间分割树,解决辅助优化器上述缺点.

5 方法扩展与当前应用领域

本节对贝叶斯优化方法扩展研究和当前主要应用领域进行总结.

5.1 贝叶斯优化方法扩展

对贝叶斯优化方法扩展可分为两类:对概率代理模型的扩展和对采集函数的扩展.

5.1.1 概率代理模型扩展

1) 高维度扩展

维数灾难(curse of dimensionality)是机器学习、数据挖掘等多领域涉及的现象.随着维度的增加,搜索空间 $\mathcal{X}$ 会以指数形式急剧增长.然而,许多高维度问题都存在低有效维度(low effective dimensionality)的性质,即:仅少数维度决定目标函数,而存在大量对目标函数影响细微或无关的维度.许多研究者应用自动相关确定技术去掉无关维度,或者通过稀疏的方法近似求解.Wang 等人通过使用随机嵌入方法把贝叶斯优化从高维度空间映射到低维度空间进行求解,该方法能够求解近十亿维的低有效维度问题(有效维度为 2)[16]. Qian 等人提出一种连续的随机嵌入方法,即在每次迭代随机乘以一个低维嵌入矩阵,使得原始维度映射到较低维度,同时放松了问题低有效维度的假设[62].Li 等人提出一种适用高维度的贝叶斯优化,该方法范化了低有效维度的假设,使用 projected-additive 的高斯过程求解高维度问题,并取得不错的效果[63].Wang 等人通过概率图模型学习变量之间潜在关系来减少不必要的评估,以解决高维度问题[64].Gardner 等人通过 MCMC 方法从高维度中发现互相独立的累加结构并进行模型选择(如高斯过程中的协方差函数),然后同时优化各部分,以达到加速优化的目的[65].Li 等人受到神经网络优化加速技术 dropout 的启发,利用该技术在每次迭代中随机丢弃固定维度,仅优化剩余维度,以达到降维的目的[66].该方法虽然简单有效但是需预先定义被抛弃维度的补全方案.

2) 多任务扩展

前面提到的贝叶斯优化仅适用于单个任务的情形,然而许多情况下希望同时优化多个相关的任务.解决这

[1] 所谓乐观是指算法在每一轮扩展叶子节点时都有可能包含最优解.

一问题的本质方法是通过一个任务提供的信息应用到其他相关任务上.Swersky 等人通过使用多输出的高斯过程[29]将贝叶斯优化扩展成多任务方法[19].该方法的核心是通过公式(42)协方差函数来捕获任务之间的关联性:

$$k((\boldsymbol{x},\boldsymbol{m}),(\boldsymbol{x}',\boldsymbol{m}'))=k_x(\boldsymbol{x},\boldsymbol{x}')\otimes k_f(\boldsymbol{m},\boldsymbol{m}'), \tag{42}$$

其中,k_x表示 $\boldsymbol{x}$ 之间的协方差函数,k_f表示任务之间的协方差函数,$\otimes$ 表示克罗内克积.

Bonilla 等人提出的多任务扩展方法是将公式(42)中 $\boldsymbol{m}$ 看作附加信息(如与任务相关的特征),从而引入其他源信息,增加任务关联性[67].而之后 Bonilla 等人又提出一种自由形式的多任务高斯过程,即公式(42)中 k_f 构成的协方差矩阵中所有元素皆为参数,通过数据学习任务之间的关联性,增加模型灵活性[68].

3） 冻融(Freeze-Thaw)扩展

在试验设计时,传统的贝叶斯优化在完全训练后才能对模型的性能进行评估,然而模型训练需要花费大量的时间,因此希望在训练模型的同时能够预先评估模型的性能.Swersky 等人提出一种冻融贝叶斯优化[69].“冻”表示挂起未完全训练的模型,“融”表示继续训练某个未完全训练的模型.该方法能在训练过程中预测模型性能,使其能挂起表现相对不好的试验并恢复表现相对好的试验.同时,该方法构造了一个非稳定性的协方差函数来预测模型的性能:

$$k_t(t,t')=\frac{\beta^{\alpha}}{(t+t'+\beta)^{\alpha}}. \tag{43}$$

并且通过基于熵搜索的采集函数选择继续训练的模型.

5.1.2 采集函数扩展

1） 约束和代价敏感性

在优化现实生活中的实际问题时,有时需要满足某些预先定义的约束.例如:食品工厂生产一种饼干,目标是使饼干具有最低的卡路里,同时需要满足大多数人的口味.实际优化问题中的这些约束有时也是黑箱的.针对这一问题,Gelbart 等人提出一种解决带黑箱约束的贝叶斯优化[70].该方法假设约束之间相互独立,并提出一种结合约束的采集函数:

$$\alpha_t(\boldsymbol{x};D_{1:t})=EI(\boldsymbol{x})\prod_{k=1}^{K}p(C_k(\boldsymbol{x})), \tag{44}$$

其中,$EI(\boldsymbol{x})$表示 EI 采集函数,$p(C_k(\boldsymbol{x}))$表示 $\boldsymbol{x}$ 满足约束 k 的概率.由于评估目标函数和约束同样需要大量的饼干制作、卡路里及口味测试过程,花费高昂的代价.因此该方法把目标函数和约束看作多任务处理,即在每次迭代仅选择一个任务(目标函数或约束)进行评估.选择任务的方法是基于熵搜索的方法,选择可以提供最多信息的任务进行评估.

如果优化过程中需要考虑时间消耗或存储量有明确预算等,并且每个 $\boldsymbol{x}$ 的评估代价存在差异,那么选择策略需要考虑评估的代价敏感性.针对这一问题,Snoek 等人提出一种具有代价敏感性的采集函数[18]:

$$\alpha_t(\boldsymbol{x};D_{1:t})=EI(\boldsymbol{x})/c(\boldsymbol{x}), \tag{45}$$

其中,$c(\boldsymbol{x})$表示 $\boldsymbol{x}$ 配置所需要的代价.公式(45)可以简单理解为单位时间的期望提升(若时间敏感),即偏向选择单位时间提升最大的 $\boldsymbol{x}$ 进行评估,从而在指定预算下得到理想解.

Kandasamy等人[71]和Marco等人[72]均考虑多精度评估情形(例如模拟的精度低但代价低,而实际试验精度高但代价高),并分别使用基于置信区间的策略和熵搜索策略自动平衡模拟和实验之间的选择,从而达到最小代价的目的.

2） 基于距离的采集函数

Marchant 等人提出一种基于距离的采集函数[14].基于距离的采集函数如下:

$$\alpha_t(\boldsymbol{x};D_{1:t})=s(\boldsymbol{x})-r\times d(\boldsymbol{x}\,|\,\boldsymbol{x}^{-}), \tag{46}$$

其中,$s(\boldsymbol{x})$表示任意的采集函数,$d(\boldsymbol{x}|\boldsymbol{x}^{-})$表示当前采样点与上一次采样点的距离,$r>0$ 表示惩罚参数.

这类采集函数可以应用于距离敏感的采样过程,如:在环境监控中,需要通过飞行器飞行到达某一位置去采样当地污染物浓度,利用基于距离的采集函数不仅保证通过少量采样找到污染物浓度最大的地区,而且保证飞

行距离最小.

3） 并行化扩展

贝叶斯优化本质是一个序贯模型.但为了加快贝叶斯优化的求解效率,可同时进行多次函数评估,即并行化扩展.Ginsbourger 等人提出一种并行化方法[73],该方法主要思想是结合使用已观测的数据集 $D_{1:t}$ 和正在观测且尚未结束的数据集 $D_{1:p}=\{(\boldsymbol{x}'_1,y'_1),(\boldsymbol{x}'_2,y'_2),...,(\boldsymbol{x}'_p,y'_p)\}$ 选择下一个评估点.当 y'_p 为常数时,该方法称为 constant liar 策略.当 y'_p 为 $\boldsymbol{x}'_p$ 在数据集 $D_{1:t}$ 上的高斯预测均值时,该方法称为 Kriging believer 策略.

Snoek 等人提出一种并行化的采集函数[18]:

$$\alpha_t(\boldsymbol{x};D_{1:t},D_{1:p})=\int\alpha_t(\boldsymbol{x};D_{1:t}\cup D'_{1:p})p(y'_{1:p}\mid D_{1:t})dy'_{1:p}\approx\frac{1}{S}\sum_{s=1}^{S}\alpha_t(\boldsymbol{x};D_{1:t}\cup D'^{(s)}_{1:p}), \tag{47}$$

其中, $D'^{(s)}_{1:p}\sim p(y'_{1:p}\mid D_{1:t})$.实验证明当采集函数为EI时,该并行采集函数具有优秀的求解效率.虽然该方法能同时评估目标函数,但是该方法的候选点仍然是顺序得到的,并不是严格的并行.严格的并行方法是一次产生多个候选点.Hutter 等人[74]提出一种能够同时产生多个候选点并同时评估的并行方法.

5.1.3 扩展方法对比

为方便相关领域研究者对当前贝叶斯优化的扩展方法有一个清晰认识,本节对上述扩展方法进行分类、总结和对比,并列出其最具代表性的扩展方法.表 6 和表 7 分别对比了概率代理模型和采集函数的扩展方法.

Table 6 Comparison of extension methods of probabilistic surrogate models

表 6 概率代理模型扩展方法对比

类别	代表性方法	特点描述及适用场景
高维度扩展	Wang 等人[16]	利用随机嵌入降维,能求解近十亿维问题,但需满足低有效维度假设
	Qian 等人[62]	利用连续随机嵌入降维,放松了低有效维度假设
	Li 等人[63]	泛化了低有效维度假设,利用在优化过程中根据最大边际似然学习分组,增加分组的计算量
	Wang 等人[64]	利用概率图模型更准确地学习分组结构,但同样引入分组计算量
	Gardner 等人[65]	利用 MCMC 学习分组并根据似然自动选择模型(如协方差函数)
	Li 等人[66]	简单有效且附加计算量少,但需要预先定义维度补充方法
多任务扩展	Swersky 等人[19]	利用多输出高斯过程代理目标函数
	Bonilla 等人[67]	引入多源信息增强任务之间关联性,适用于具有多源任务相关信息的情形
	Bonilla 等人[68]	通过数据学习任务关联性,增加模型灵活性,由于参数增加,因此需保证足够数据量
冻融扩展	Swersky 等人[69]	能在训练过程中预测模型性能,挂起表现不好的试验,恢复表现好的试验,从而加快优化进程,适用于能随时暂停和恢复训练且完整训练极其耗时的问题

Table 7 Comparison of extension methods of acquisition functions

表 7 采集函数扩展方法对比

类别	代表性方法	特点描述及适用场景
约束和代价敏感性	Gelbart 等人[70]	能处理带黑箱约束的问题,但需假设各个约束相互独立
	Snoek 等人[18]	考虑代价敏感性,适用于候选点评估代价不同的问题
	Kandasamy 等人[71]	适用于多精度评估问题,基于置信区间策略选择不同精度评估
	Marco 等人[72]	适用于多精度评估问题,基于熵搜索策略选择不同精度评估
距离敏感	Marchant 等人[14]	适用于对移动距离敏感的问题,如利用无人机监控交通或环境
并行化	Ginsbourger 等人[73]	为未完成的观测分配假的观测值(如常量或均值),用于选择之后的评估点,但假的观测会影响候选点选择结果
	Snoek 等人[18]	考虑了未完成观测的所有可能情况,用采样方式近似积分,以精确地选择之后地评估点,但引入积分近似的计算量,且不是严格上的并行
	Hutter 等人[74]	能同时产生多个候选点并同时评估

5.2 当前应用领域

作为优化复杂黑箱问题的有效手段,贝叶斯优化已被应用于许多领域.本节将详细地总结其当前应用领域.

1) A/B 测试、游戏与材料设计

Google 和 Microsoft 等公司在广告与网页优化设计方面[2,3,4]应用了贝叶斯优化.解决的问题是,在一定查询

预算的前提下,如何择优选择用户进行查询[I],帮助设计和改善产品.利用广告、网页、应用程序途径等得到的用用户反馈,开发者可通过贝叶斯优化对产品的配置进行优化调整.Khajah 等人利用贝叶斯优化设计出最大化用户参与度的游戏[5].他们通过调整游戏中的设置,如:敌人个数、出现频率、开枪次数等,来控制游戏难度,将玩家参与游戏的时间作为反馈,优化出用户参与度最高的游戏配置.Frazier 等人应用贝叶斯优化进行材料设计,选取合适的化学结构、组成成分和处理条件等构造理想的材料[6].

2) 推荐系统

Google 和 Microsoft 等公司应用贝叶斯优化技术,根据订阅者订阅的网站、视频、音乐等方面内容为订阅者推荐相关的新闻文章[7,8].A/B 测试与游戏设计每次迭代只能给出一个网页或者游戏配置,然而推荐系统可以一次为任意订阅者推荐多个新闻或者商品.

3) 机器人学、嵌入式系统及系统设计

对两足或多足机器人的步态优化十分有挑战性.Lizotte 等人应用贝叶斯优化解决传统步态优化方法容易陷入局部最优和需要大量评估的缺点[11].该方法采用高斯过程作为概率代理模型,采用 PI 采集函数实现了更快、更平稳、评估次数更少的机器人步伐评估过程.Martinez-Cantin 等人提出一种在有限视野和局部观测下的基于模拟的主动策略学习算法(高斯过程代理模型,EI 采集函数),应用于机器人导航和不确定性地点探索[12].Schneider 讨论了嵌入式学习系统的挑战和贝叶斯优化应用到嵌入式学习系统的发展前景[13].Akrour 等人利用局部环境的贝叶斯优化在高维度空间(70 维)中控制机器人臂运动[75].Torun 等人提出一种两阶段贝叶斯优化方法(第一阶段注重不确定性区域探索,第二阶段根据当前探索区域寻找最优)优化集成系统设计[76].

4) 环境监控与传感器网络

传感器设备用于测量速度、温度、湿度、空气质量、污染物含量等环境指标.由于不能在所有区域布置传感器,再加上噪声的干扰,传感器测量的数据常常存在不确定性.此外,激活传感器设备进行环境感知都会消耗能量,如:电量和传输流量.Srinivas 等人使用高斯过程代理的贝叶斯优化,通过仅激活少量的传感器,便可找到室内温度极值位置或高速公路上最堵位置[23].Garnett 等人使用贝叶斯优化选择最优传感器子集,使其根据这些子集得到最优的预测效果[24].Marchant 等人把贝叶斯优化扩展到环境监控中,利用可移动机器人在环境中进行主动采样,得到对周围环境的精确感知[14].Morere 等人结合贝叶斯优化和部分观测的马尔科夫决策过程,以优化无人机采样策略监测周围环境[77].Colopy 等人利用贝叶斯优化调整基于个体的个性化监测模型,以个性化地监控病人生命体征[78].Candelieri 等人利用贝叶斯优化优化控制给水管网系统中的泵,以达到在少量能量消耗的情况下得到理想的泵调度方案的目的[79].

5) 偏好学习与交互界面

在处理计算机图形与动画领域中的问题时,通常需要专业人员手动调整大量棘手的参数.例如:构造烟雾场景的粒子系统,需要调整速度、半径、涡环大小、长度尺度、旋度噪音等参数.通常情况下,这些参数十分复杂,非专业人员难以理解.Brochu 等人提出一种使用贝叶斯优化的迭代选择方法.该方法在处理图片时不需要专业人员手动调参,只需在每次迭代从生成的两张对比图片[II]中选取与目标更像的图片作为反馈[III],不需要用户理解复杂参数的具体含义.该方法通过返回的对比偏好信息更新代理模型,并根据完全随机、EI 等策略生成下一次迭代的两张对比图片,直到找到满足需求的目标图片[9,10].

6) 自动算法配置

构造一个优秀的算法通常需要经过大量的参数调节试验.若算法的参数调节都需要人工干预,将花费大量的时间和人力,甚至做无用功.因此,自动算法配置十分必要.这样不仅能减少人工干预,使得人们能更专注于新模型构建等高层次问题,还能减少大量的训练时间.相比人工经验或穷举,优化算法会自动选择合适的参数配置进行训练验证.贝叶斯优化能够胜任这类问题,并已取得了令人瞩目的成果.Bergstra 等人应用贝叶斯优化自动

I 这里的查询是指一个用户对某版本的产品进行评测,返回点击率或其他测度.

II 两张对比图片具有不同的参数配置.

III 此时用户知道最终想要的图片效果.

地调整神经网络和深度信念网络中的超参数[17].Snoek 等人应用贝叶斯优化自动调整卷积神经网络中的超参数[18,19].Mahendran 等人提出一种基于贝叶斯优化的自适应马尔科夫链蒙特卡洛算法[20].Thornton 等人应用贝叶斯优化提出一种针对分类算法的自动模型选择和超参数调节的方法:Auto-WEKA[21].Zhang 等人使用贝叶斯优化对卷积神经网络中的参数进行调整解决目标识别问题[15].Wang 等人通过贝叶斯优化调整混合整数规划求解器的参数来提升求解器的效率[16]. Klein 等人提出一种快速贝叶斯优化方法,能够调节大规模数据集上的机器学习算法的超参数[80].Xia 等人应用贝叶斯优化调节决策树中超参数,提高信用评价精度[81].

7) 自然语言与文本处理

Wang 等人使用贝叶斯优化对文本进行术语提取(Term Extraction)[61].Yogatama 等人利用贝叶斯优化为不同类问题选择合适的文本表示.其实验证明,该方法能使优化后的线性模型与未优化的复杂模型在主题分类问题上具有可比的效率[82].

8) 生物、化学及晶体学

贝叶斯优化同样可以胜任在生物、化学及晶体学等领域中的高代价优化任务.Carr 等人应用贝叶斯优化技术在晶体表面上寻找分子最稳定的吸附位置[83].Krivák 等人用贝叶斯优化提升配体成键位置的预测质量[84].Tanaka 等人利用贝叶斯优化进行全基因组选择,能够在少量的模拟代价下找到较理想基因型[85].在脑年龄分类预测任务中,Lancaster 等人利用贝叶斯优化调节对神经影像预处理时所采用重采样技术的参数,进而达到提高分类精度的目的[86].

9) 迁移学习

Ruder 等人在迁移学习过程中,利用贝叶斯优化技术从多源或多领域数据中自动地选择有效数据作为训练集,以达到增强模型能力的目的,且与具体学习模型无关[87].

6 问题与挑战

前面详细地介绍了贝叶斯优化的研究现状.然而,随着大数据应用的发展,待优化目标的规模和复杂程度将增加.作为处理评估代价大的复杂黑箱问题的有效解决方法,贝叶斯优化在未来发展中将面临下列问题与挑战.

一、 实时性和自适应性

贝叶斯优化每次迭代需要对概率代理模型进行更新,当问题维度高或存在大量历史数据时,更新概率模型需要高昂的计算量,尤其不能满足对实时性要求高的实际任务.针对该问题,研究者已经提出了一些解决策略.1)降维映射,见 5.1.1 节.当贝叶斯优化处理高维度问题时,需要从高维度空间映射到低维度空间进行优化,虽然该方法加快了求解效率,但是需要假设问题存在低有效维度的性质.2)近似方法,见 4.1 节.当模型的先验不为共轭先验时,需要使用变分贝叶斯近似推断或蒙特卡洛采样方法得到模型近似后验分布.当使用高斯过程代理目标函数时,精确推断需要 $O(t^3)$的时间复杂度,可使用 Cholesky 分解、SPGP、SSGP 等方法对高斯过程进行近似推断.虽然这些近似方法能够加快求解效率,但是具有求解精度不足的缺点.3)并行化,见 5.1.2 节.通过对贝叶斯优化进行并行化扩展,能够同时评估多次目标函数,加快求解效率.该策略选择评估点时根据部分未完成评估的采样点返回的虚拟观测值,而不是真实观测值,会在一定程度上影响求解精度.4)时间敏感性(5.1.2 节).时间敏感性主动选择策略能够选择单位时间期望提升最大的点进行评估.但该方法在相同迭代预算下与传统方法相比存在精度差异.在提高贝叶斯优化求解效率时,难点在于如何解决精度和计算开销之间的平衡关系.

此外,贝叶斯优化在处理优化目标动态变化的问题时,应该具有自适应调整能力.在已有规划解的基础上,针对问题变化,动态调整现有策略,而不需要推倒重来,从头计算.例如:在交通领域中,当车辆前方发生不可预测的事件(如车祸)造成拥堵时,需要优化程序能够自适应地、增量地调整规划路线.

二、 分布式

随着数据量的增加,复杂应用很难在一台终端上高效执行.因此,贝叶斯优化还需要具有分布式处理数据的能力.贝叶斯优化的分布式扩展应具有以下特点:

1)负载均衡.能够有效地利用计算资源,避免资源过于集中和浪费.2)具有高效的计算效率.目前贝叶斯优化

并行技术是为了加快其求解效率,同时进行多次函数评估,本质上是对采集函数的并行化扩展(见 5.1.2 节).该方法仍存在集中环节,即集中回收评估点返回的观测值集合,然后整合更新概率模型决策候选点集合.3)高容错性和强健壮性.分布式计算中一个任务往往存在多个备份,一个备份所在终端失效后,其余备份仍可继续执行,从而实现任务的健壮性.,与之不同的是,贝叶斯优化过程所要求的高容错性和强健壮性应能有效处理没有备份的任务,根据需要动态的进行优化策略调整.例如:在无人机对抗情景中,将每个无人机看作节点,这些无人机基于自组织、不可靠的通讯网进行协同作战.当一架无人机被击落时,该小组应能动态调整队形,继续执行作战任务.这种去中心化的优化策略可以避免出现击毁中心机使整个小组瘫痪的情况.4)多策略分布式协同求解.贝叶斯优化的分布式扩展可同时存在多个不同的策略(不同的概率模型和采集函数),并像深度学习中的对抗网络一样,各个策略相互促进、相互影响,从而达到理想的学习效果.然而,对贝叶斯优化分布式扩展的难点在于分布式概率代理模型和采集函数的构建,并且需要处理各个分散节点之间的信息交互问题.

三、 多目标

贝叶斯优化的多任务扩展能够处理多个相关任务,根据相关性,将一个任务的信息应用到其他相关任务上,从而达到迁移学习的目的.例如:5.1.1 节中 Swersky 等人使用高斯过程同时处理多个相关的超参数优化任务,为每一个任务得到最优的超参数配置,从而使系统性能最大化.该方法的优化目标是最优化所有任务的平均性能.但在实际应用中,许多问题需要同时优化多个目标,这些目标可能会存在“冲突”关系.例如:在智能交通应用中,既要规划出最短路径又要尽量多的收集未知区域的道路情况,但这两个目标很难同时满足.5.1.2 节中介绍的约束扩展方法将两个目标中的一个作为优化目标,另一个作为约束处理.当目标间存在冲突时,不存在绝对最优解,只存在有效解集合.当把多目标转换成带约束的单目标优化时,求得的优化解仅是单目标的最优解,忽略了转化为约束的目标的重要程度.Tesch 等人提出一种面向多目标的贝叶斯优化方法,尽管该方法能够得到帕累托集,但忽略了目标之间的依赖关系[88].多目标贝叶斯优化的难点在于处理多个目标之间的关系.为了保证所有目标的重要性并利用目标之间的依赖关系,贝叶斯优化在求解多目标问题时可考虑同时拥有多个概率代理模型和采集函数,在优化过程中,这些概率模型和采集函数相互促进,相互影响,达到优化学习的目的.

四、 模型选择问题

模型选择一直是贝叶斯方法面临的棘手问题.贝叶斯优化涉及的模型选择有:观测模型选择、(非)参数模型先验选择以及超参数先验选择.观测模型需根据领域知识指导选择.合理的观测模型需对错误假设具有鲁棒性,即当真实数据与模型假设不相符时其仍具有良好的表现.不同问题具有不同的性质,因而具有不同的先验形式.例如:在监测城市道路状况时,由于人们有早出晚归的习惯,通常道路状况会出现早晚高峰周期性的表现,因此可以选择存在周期性质的协方差函数构造先验模型.然而极端环境监测问题,不具备这样的周期性质,因此需要选择其它合适的先验模型.当使用贝叶斯方法估计超参数时需要选择合理的超参数先验,增加超参数估计精度,提升模型预测准确率.

在贝叶斯优化中,选择合适的概率代理模型甚至比采集函数的选择还要关键.在一些领域中,如制药和传染病控制,需要更加谨慎地选择合适的模型,提高概率模型预测的准确度,减少评估过程的代价.尽管目前存在一些模型选择的方法[29],但这些方法都不具有通用性,仍需要针对具体问题具体分析,运用领域专家的经验知识指导模型的选择.在贝叶斯优化研究和应用领域,如何针对具体问题选择合适的概率代理模型仍是具有很大挑战性的问题.

7 总结

作为求解非凸、多峰、评估代价高昂、黑箱的复杂优化问题的有效解决方案,贝叶斯优化近年来在多领域获得了广泛关注.本文综述了贝叶斯优化的研究现状.首先从其优化框架和优化原理入手,详细分析其优势与劣势,以帮助相关领域研究者深入理解贝叶斯优化;然后从模型选择的角度介绍了贝叶斯优化两个核心部分:概率代理模型和采集函数,旨在建模求解复杂优化问题进行模型选择时提供参考;其次介绍了贝叶斯优化涉及的近似与优化技术,深入技术细节;最后总结了贝叶斯优化的方法扩展和当前主要应用领域;同时,本文也关注随

着待优化目标的规模和复杂程度的增加,贝叶斯优化将面临实时性和自适应性、分布式、多目标以及模型选择等问题与挑战.此外,相比于其它优化技术,贝叶斯优化还存在一些局限性.本文通过对贝叶斯优化的详细分析和讨论,希望为相关领域的研究者予以帮助.

References:

[1] Shahriari B, Swersky K, Wang Z, Adams RP, Freitas ND. Taking the human out of the loop: a review of Bayesian optimization. In: Proceedings of the IEEE, 2016, 104(1):148-175.

[2] Kohavi R, Longbotham R, Dan S, Henne RM. Controlled experiments on the web: survey and practical guide. Data Mining and Knowledge Discovery, 2009, 18(1):140-181.

[3] Scott SL. A modern Bayesian look at the multi-armed bandit. Applied Stochastic Models in Business and Industry, 2010, 26(6):639-658.

[4] Chapelle O, Li L. An empirical evaluation of Thompson sampling. In: Advances in Neural Information Processing Systems, 2011:2249-2257.

[5] Khajah MM, Roads BD, Lindsey RV, Liu YE, Mozer MC. Designing engaging games using Bayesian optimization. In: Proceedings of the ACM Conference on Human Factors in Computing Systems, 2016:5571-5582.

[6] Frazier PI, Wang J. Bayesian optimization for materials design. Mathematics, 2015.

[7] Li L, Chu W, Langford J, Schapire RE. A contextual-bandit approach to personalized news article recommendation. In: Proceedings of the International Conference on World Wide Web, 2010:661-670.

[8] Vanchinathan HP, Nikolic I, Bona FD, Krause A. Explore-exploit in top-n recommender systems via Gaussian processes. In: Proceedings of the ACM Conference on Recommender Systems, 2014:31.

[9] Brochu E, Brochu T, Freitas ND. A Bayesian interactive optimization approach to procedural animation design. In: Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation, 2010:103-112.

[10] Brochu E, Cora VM, Freitas ND. A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. Computer Science, 2010.

[11] Lizotte D, Wang T, Bowling M, Schuurmans D. Automatic gait optimization with Gaussian process regression. In: Proceedings of the International Joint Conference on Artifical Intelligence, 2007:944-949.

[12] Martinez-Cantin R, Freitas ND, Doucet A, Castellanos JA. Active policy learning for robot planning and exploration under uncertainty. In: Proceedings of Robotics: Science and Systems III, 2007:321-328.

[13] Schneider J. Bayesian optimization and embedded learning systems. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016:413-413.

[14] Marchant R, Ramos F. Bayesian optimisation for intelligent environmental monitoring. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, 2012:2242-2249.

[15] Zhang Y, Sohn K, Villegas R, Pan G, Lee H. Improving object detection with deep convolutional networks via Bayesian optimization and structured prediction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015:132-132.

[16] Wang Z, Zoghi M, Hutter F, Matheson D, Freitas ND. Bayesian optimization in a high dimensions via random embeddings. In: Proceedings of the International Joint Conference on Artificial Intelligence, 2013.

[17] Bergstra J, Bardenet R, Bengio Y, Kégl B. Algorithms for hyper-parameter optimization. In: Advances in Neural Information Processing Systems, 2011, 24(24):2546-2554.

[18] Snoek J, Larochelle H, Adams RP. Practical Bayesian optimization of machine learning algorithms. In: Advances in Neural Information Processing Systems, 2012, 4:2951-2959.

[19] Swersky K, Snoek J, Adams RP. Multi-task Bayesian optimization. In: Advances in Neural Information Processing Systems, 2013:2004-2012.

[20] Mahendran N, Wang Z, Hamze F, Freitas ND. Adaptive MCMC with Bayesian optimization. In: Proceedings of the International Conference on Artificial Intelligence and Statistics, 2010.

[21] Thornton C, Hutter F, Hoos HH, Leyton-Brown K. Auto-WEKA: combined selection and hyperparameter optimization of classification algorithms. Computer Science, 2013:847-855.

[22] Hoffman MW, Shahriari B and Freitas ND. On correlation and budget constraints in model-based bandit optimization with application to automatic machine learning. In: Proceedings of the International Conference on Artificial Intelligence and Statistics, 2014:365–374.

[23] Srinivas N, Krause A, Kakade SM, Seeger M. Gaussian process optimization in the bandit setting: no regret and experimental design. In: Proceedings of the International Conference on Machine Learning, 2010.

[24] Garnett R, Osborne MA, Roberts SJ. Bayesian optimization for sensor set selection. In: Proceedings of the International Conference on Information Processing in Sensor Networks, 2010:209-219.

[25] Ghahramani Z. Probabilistic machine learning and artificial intelligence. Nature, 2015, 521:452-459.

[26] Jones DR, Schonlau M, Welch WJ. Efficient global optimization of expensive black-box functions. Journal of Global Optimization, 1998, 13(4):455-492.

[27] Nelder JA, Baker RJ. Generalized linear models. Journal of the Royal Statistical Society, 1972, 135(3):370-384.

[28] Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. In: Advances in Neural Information Processing Systems, 2014, 4:3104-3112.

[29] Rasmussen CE, Williams CKI. Gaussian processes for machine learning. The MIT Press, 2006. ISBN 0-262-18253-X.

[30] Lu C, Tang X. Surpassing human-level face verification performance on LFW with GaussianFace. Computer Science, 2014.

[31] Neal RM. Bayesian learning for neural networks. [Ph.D. Thesis]. University of Toronto, 1996.

[32] Paciorek CJ, Schervish MJ. Nonstationary covariance functions for Gaussian process regression. In: Advances in Neural Information Processing Systems, 2003, 16:273-280.

[33] Hutter F, Hoos HH, Leyton-Brown K. Sequential model-based optimization for general algorithm configuration. In: Proceedings of the Conference on Learning and Intelligent Optimization, 2011:507-523.

[34] Watson, GN. A treatise on the theory of Bessel functions. 2nd ed. Cambridge University Press, 1966.

[35] Breiman L. Random forests. Machine Learning, 2001, 45(1):5-32.

[36] Zhang Y, Chan W, Jaitly N. Very deep convolutional networks for end-to-end speech recognition. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing. 2017.

[37] Karpathy A, Fei-Fei L. Deep visual-semantic alignments for generating image descriptions. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2014, 39(4):664-676.

[38] Snoek J, Rippel O, Swersky K, Kiros R, Satish N, Sundaram N, Patwary MMA, Prabhat , Adams RP. Scalable Bayesian optimization using deep neural networks. Statistics, 2015:1861-1869.

[39] Springenberg JT, Klein A, Falkner S, Hutter F. Bayesian optimization with robust Bayesian neural networks. In: Advances in Neural Information Processing Systems, 2016

[40] Kushner HJ. A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise. Journal of Fluids Engineering, 1963, 86(1).

[41] Jones DR. A taxonomy of global optimization methods based on response surfaces. Journal of Global Optimization, 2001, 21(4):345-383.

[42] Mockus J, Tiesis V, Zilinskas A. The application of Bayesian methods for seeking the extremum. Towards Global Optimisation 2. 1978:117-129.

[43] Lizotte DJ . Practical Bayesian optimization. [Ph.D. Thesis]. University of Alberta, 2008.

[44] Lai TL, Robbins H. Asymptotically efficient adaptive allocation rules. Advances in Applied Mathematics, 1985, 6(1):4-22.

[45] Thompson WR. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika, 1933, 25(3-4):285-294.

[46] Shahriari B, Wang Z, Hoffman MW, Bouchard-Côté. An entropy search portfolio for Bayesian optimization. In: Proceedings of the Conference on Neural Information Processing Systems: Workshop on Bayesian Optimization in Academia and Industry, 2014.

[47] Lázaro-Gredilla M, QuiñOnero-Candela J, Rasmussen CE, Figueiras-Vidal AR. Sparse spectrum Gaussian process regression. Journal of Machine Learning Research, 2010, 11(9):1865-1881.

[48] Villemonteix J, Vazquez E, Walter E. An informational approach to the global optimization of expensive-to-evaluate functions. Journal of Global Optimization, 2009, 44(4):509-534.

[49] Hernández lobato JM, Hoffman MW, Ghahramani Z. Predictive entropy search for efficient global optimization of black-box functions. In: Proceedings of the Conference on Neural Information Processing Systems: Workshop on Bayesian Optimization in Academia and Industry, 2014.

[50] Brochu E, Hoffman M, Freitas ND. Portfolio allocation for Bayesian optimization. In: Proceedings of the Conference on Uncertainty in Artificial Intelligence, 2011.

[51] Tzikas DG, Likas CL, Galatsanos NP. The variational approximation for Bayesian inference. IEEE Signal Processing Magazine, 2008, 25(6):131-146.

[52] Seeger M, Williams CKI, Lawrence ND. Fast forward selection to speed up sparse Gaussian process regression. In: Proceedings of the Conference on Artificial Intelligence and Statistics, 2003.

[53] Snelson E, Ghahramani Z. Sparse Gaussian process using pseudo-inputs. In: Advances in Neural Information Processing Systems, 2006, 18(1):1257-1264.

[54] Martinez-Cantin R. BayesOpt: a Bayesian optimization library for nonlinear optimization, experimental design and bandits. Journal of Machine Learning Research, 2014, 15:3735-3739.

[55] Osborne MA, Garnett R, Roberts SJ. Gaussian processes for global optimization. In: Proceedings of the International Conference on Learning and Intelligent Optimization, 2009.

[56] Rasmussen CE, Ghahramani, Z. Bayesian Monte Carlo. In: Advances in Neural Information Processing Systems. 2002.

[57] Osborne MA, Roberts SJ, Rogers A, Ramchurn SD, Jennings NR. Towards real-time information processing of sensor network data using computationally efficient multi-output Gaussian processes. In: Proceedings of the International Conference on Information Processing in Sensor Networks, 2008:109–120

[58] Bardenet R, Kégl B. Surrogating the surrogate: accelerating Gaussian-process-based global optimization with a mixture cross-entropy algorithm. In: Proceedings of the International Conference on Machine Learning, 2010.

[59] Jones DR, Perttunen CD, Stuckman BE. Lipschitzian optimization without the Lipschitz constant. Journal of Optimization Theory and Applications, 1993, 79(1):157-181.

[60] Hansen N, Ostermeier A. Completely derandomized self-adaptation in evolution strategies. IEEE Trans. on Evolutionary Computation, 2001, 9(2):159-195.

[61] Wang Z, Shakibi B, Jin L, Freitas ND. Bayesian multi-scale optimistic optimization. In: Proceedings of the International Conference on Artificial Intelligence and Statistics, 2014:1005-1014.

[62] Qian H, Hu YQ, Yu Y. Derivative-free optimization of high-dimensional non-convex functions by sequential random embeddings. In: Proceedings of the International Joint Conference on Artificial Intelligence. 2016.

[63] Li CL, Kandasamy K, Poczos B, Schneider J. High dimensional Bayesian optimization via restricted projection pursuit models. In: Proceedings of the International Conference on Artificial Intelligence and Statistics, 2016

[64] Wang Z, Li C, Jegelka S, Kohli P. Batched high-dimensional Bayesian optimization via structural kernel learning. In: Proceedings of the International Conference on Machine Learning, 2017.

[65] Gardner JR, Guo C, Weinberger KQ, Garnett R, Grosse R. Discovering and exploiting additive structure for Bayesian optimization. In: Proceedings of the International Conference on Artificial Intelligence and Statistics, 2017.

[66] Li C, Gupta S, Rana S, Nguyen V, Venkatesh S, Shilton A. High dimensional Bayesian optimization using dropout. In: Proceedings of the International Joint Conference on Artificial Intelligence. 2017.

[67] Bonilla EV, Agakov FV, Williams CKI. Kernel multi-task learning using task-specific features. In: Proceedings of the International Conference on Artificial Intelligence and Statistics, 2007.

[68] Bonilla EV, Chai KMA, Williams CKI. Multi-task Gaussian process prediction. In: Advances in Neural Information Processing Systems, 2007.

[69] Swersky K, Snoek J, Adams RP. Freeze-Thaw Bayesian optimization. Eprint Arxiv, 2014.

[70] Gelbart MA, Snoek J, Adams RP. Bayesian optimization with unknown constraints. Computer Science, 2014.

[71] Kandasamy K, Dasarathy G, Oliva J, Schneider J, Poczos B. Gaussian process bandit optimisation with multi-fidelity evaluations. In: Advances in Neural Information Processing Systems. 2016.

[72] Marco A, Berkenkamp F, Hennig P, Schoellig AP, Krause A, Schaal S, Trimpe S. Virtual vs. real: trading off simulations and physical experiments in reinforcement learning with Bayesian optimization. In: Proceedings of the International Conference on Robotics and Automation. 2017.

[73] Ginsbourger D, Riche RL, Carraro L. Kriging is well-suited to parallelize optimization. Computational Intelligence in Expensive Optimization Problems, 2010:131-162.

[74] Hutter F, Hoos HH, Leyton-Brown K. Parallel algorithm configuration. In: Proceedings of the International Conference on Learning and Intelligent Optimization, 2012:55-70.

[75] Akrour R, Sorokin D, Peters J, Neumann G. Local Bayesian optimization of motor skills. In: Proceedings of the International Conference on Machine Learning. 2017.

[76] Torun HM, Swaminathan M, Davis AK, Bellaredj MLF. A global Bayesian optimization algorithm and its application to integrated system design. IEEE Trans. on Very Large Scale Integration Systems, 2018:1-11.

[77] Morere P, Marchant R, Ramos F. Sequential Bayesian optimization as a POMDP for environment monitoring with UAVs. In: Proceedings of the International Conference on Robotics and Automation. 2017:6381-6388.

[78] Colopy GW, Roberts SJ, Clifton DA. Bayesian optimization of personalized models for patient vital-sign monitoring. IEEE Journal of Biomedical and Health Informatics, 2018, 22(2):301.

[79] Candelieri A, Perego R, Archetti F. Bayesian optimization of pump operations in water distribution systems. Journal of Global Optimization. 2018.

[80] Klein A, Falkner S, Bartels S, Henning P, Hutter F. Fast Bayesian optimization of machine learning hyperparameters on large datasets. In: Proceedings of the International Conference on Artificial Intelligence and Statistics. 2017.

[81] Xia Y, Liu C, Li YY, Liu N. A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. Expert Systems with Applications, 2017, 78:225-241.

[82] Yogatama D, Kong L, Smith NA. Bayesian optimization of text representations. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2015: 2100–2105.

[83] Carr S, Garnett R, Lo C. BASC: applying Bayesian optimization to the search for global minima on potential energy surfaces. In: Proceedings of the International Conference on Machine Learning, 2016.

[84] Krivák R, Hoksza D, Škoda P. Improving quality of ligand-binding site prediction with Bayesian optimization. In: Proceedings of the International Conference on Bioinformatics and Biomedicine. 2017:2278-2279.

[85] Tanaka R, Iwata H. Bayesian optimization for genomic selection: a method for discovering the best genotype among a large number of candidates. Theoretical and Applied Genetics, 2017, 131(1):1-13.

[86] Lancaster J, Lorenz R, Leech R, Cole JH. Bayesian optimization for neuroimaging pre-processing in brain age classification and prediction. Frontiers in Aging Neuroscience, 2018.

[87] Ruder S, Plank B. Learning to select data for transfer learning with Bayesian optimization. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2017:372-382.

[88] Tesch M, Schneider J, Choset H. Expensive multiobjective optimization and validation with a robotics application. In: Proceedings of the Conference on Neural Information Processing Systems: Workshop on Bayesian Optimization and Decision Making, 2012.