Selection of output mode for Bayesian optimization in noisy problems

Chenxi Li

2024 年 3 月 12 日

本文研究的是贝叶斯优化在有噪声问题中如何选择结果输出的方式。在无噪声的问题中，通常会直接使用所有观测值中的最优值对应点作为优化的输出结果。但在有噪声的问题中，由于所有观测值都存在噪声，这种结果的输出方式就可能不太妥当。考虑到代理模型能够提供搜索空间内任一点的预测均值，这使得我们可以选择使用预测性的输出方式。例如输出所有观测点中预测均值最优的点，或是输出整个搜索空间内预测均值最小的点。预测性的输出方式在高噪声的问题中可能会比直接输出观测值更加合适，但输出方式的选择并不仅仅只与噪声的大小有关。目标函数的响应面性质，搜索空间的维度等因素同样也会造成影响。本文使用 BBOB 函数组对带噪声的贝叶斯优化问题做了一些数值实验，并希望能够为输出方式的选择提供一些有指导性的建议。

1 Introduction

贝叶斯优化是黑盒优化领域中较为先进的优化框架，其特点在于能够以较少的目标函数评估次数来尽可能地获得接近全局最优值的解，因而被广泛使用在药物研发，神经网络参数调节等目标函数评估代价昂贵的优化问题上。该算法是一种基于模型的序贯优化 [1]，在每一步迭代时，贝叶斯优化都会根据当前已掌握的信息来制定平衡探索和开发的策略用以选择下一个评估点。具体来说，贝叶斯优化主要由代理模型和采集函数两部分组成。其中代理模型一般使用高斯过程模型，根据已有的观测点及其观测值信息来对目标函数进行建模。而后采集函数可以利用建立的模型去选择下一个观测点的位置，从而实现在较少的观测次数下得到尽量好的优化结果。

贝叶斯优化已经被验证在低维，无噪声的昂贵黑箱优化问题中有良好的优化效果。当目标函数存在噪声时，贝叶斯优化同样能够发挥作用，但其算法的设置上可能需要进行一定的改动。一个最直观的问题在于，当目标函数存在噪声时，观测点对应的观测值将不再是真实值。这个问题将会从两个方面影响贝叶斯优化的优化效果：一方面在于采集函数。目前常用的采集函数是基于最优值期

望提升的 EI，即通过计算某一点相对当前最优值的提升期望大小来选择期望最大的点作为下一个观测点。当噪声较大时，当前最优值可能会受到严重干扰，从而影响 EI 的选点策略。另一方面，噪声同样会影响到我们最终的输出结果，因为噪声情况下的观测值最小并不一定准确。解决这些问题的一个直观方案是用代理模型的预测均值最优值来代替观测最优值。这里预测均值同样也不是真实值，只是代理模型基于已有信息对于目标函数的推断。但在噪声较大的情况下，这样的预测值往往会比直接的观测值更加可信。本文主要想要探究的就是在什么样的情况下我们更应该相信观测值，什么样的条件下更应该相信预测值。当下似乎并没有文章对此作出详细的讨论，因此本文希望能够通过实验深入探究这一问题。为了使得实验结论更具代表性，本文的测试问题选择使用“BBOB”这一成熟的测试基准环境，它包含了 24 个无噪声的测试函数。我们将其修改为带噪声的形式作为测试的基准问题。

本文的大致分为以下几个部分：第一部分是对于贝叶斯优化的详细介绍，包括代理模型和采集函数。第二部分是有关 BBOB 基准测试函数的介绍以及相关的实验设定。第三部分我们先探讨了在无噪声的环境下使用观测值和使用预测值作为结果输出的优化效果差异。第四部分为带噪声的实验。最后是全文的结论总结。

2 贝叶斯优化

贝叶斯优化框架主要由两部分组成：代理模型和采集函数，用来解决复杂黑盒函数的全局优化问题。具体来说，记 $f(x)$ 为未知的目标黑盒函数，考虑优化问题：

$$x^* = \arg\max_{x \in \mathcal{X}} f(x) \tag{1}$$

其中 $\mathcal{X}$ 为搜索空间。目标函数 $f(x)$ 无需拥有显示表达式，只需满足对于任意的 $x_0 \in X$, 均可以计算出对应的目标函数值 $f(x_0)$。主流的贝叶斯优化框架均采用高斯过程 (见 2.1 节) 作为代理模型，用于构建目标函数的后验分布，具有较强的灵活性。利用高斯过程建模之后，对于搜索空间的任何一点，高斯过程模型均能返回该点对应的预测均值以及其不确定性。利用其返回的预测均值和不确定性，我们就能设计采集函数（见 2.2 节）来权衡开发与探索，以此来选择下一个观测点的位置。贝叶斯优化的具体算法流程见 **Algorithm1**。

值得注意的是，**Algorithm1** 给出的是无噪声条件下标准的贝叶斯优化算法流程。对于有噪声的问题，我们不一定会将观测值最小对应点作为输出结果。我们可能会考虑以下两种预测性的结果输出方式：输出以观测点中预测均值最小的点或代理模型在整个搜索空间上预测均值最小的点。我们会在之后的实验部分分析对比上述三种输出方式的优化表现。

Algorithm 1: Bayesian optimization

Input: number of initial point n_0, number of max iteration n

Randomly get initial data $D_{1:n_0}$ and update GP model;

for $t = 1, 2, ...n$ **do**

  Find x_t by optimizing the acquisition function over the GP $x_t = \arg\max_x u(x|D_{1:n_0+n})$;

  Sample the objective function: $y_t = f(x_t) + \epsilon_t$;

  Augment the data $D_{1:n_0+n} = D_{1:n_0+n-1}, (x_t, y_t)$ and update GP model;

end

Result: $(x_{t^*}, y_{t^*}) \in D_{1:n_0+n}$, $t^* = \arg\max y_{1:n_0+n}$

2.1 高斯过程

高斯过程是一种非参数模型，由均值函数 $m(x) : \mathcal{X} \to \mathbb{R}$ 以及正定核函数（也可视为协方差函数）$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ 组成。对于搜索空间内任何有限点集 $x_{1:n}$，定义 $f_i := f(x_i)$ 为 x_i 处对应的目标函数值，$y_i := f(x_i) + \epsilon_i$ 为 x_i 处对应的带噪声的观测值。在高斯过程模型中，我们假设 $\mathbf{f} := f_{1:n}$ 服从联合高斯分布，且在给定 $\mathbf{f}$ 的情况下 $\mathbf{y} := y_{1:n}$ 服从正态分布，具体数学表达形式如下：

$$\mathbf{f} \mid \mathcal{X} \sim \mathcal{N}(\mathbf{m}, \mathbf{K}) \tag{2}$$

$$\mathbf{y} \mid \mathbf{f}, \sigma^2 \sim \mathcal{N}(\mathbf{f}, \sigma^2\mathbf{I}]) \tag{3}$$

该模型的特点在于我们将目标函数 $f(x)$ 视为随机变量，equation(2) 代表了变量 $f(x)$ 的先验分布。其中 $\mathbf{m}$ 代表先验的均值函数,$K_{i,j} := k(x_i, x_j)$ 为协方差矩阵。得到观测数据 $\mathcal{D}_n = (x_i, y_i)_{i=1}^n$ 后，$f(x)$ 在给定观测数据 $\mathcal{D}_n$ 的条件下，对于新的待预测值 $f(x^*)$, 由高斯过程的性质，我们有如下联合分布：

$$\begin{bmatrix} \mathbf{y} \\ f(x^*) \end{bmatrix} \sim \mathcal{N}\left(\mathbf{m}, \begin{bmatrix} \mathbf{K} + \sigma^2\mathbf{I} & \mathbf{k}(x^*) \\ \mathbf{k}(x^*)^T & k(x^*, x^*) \end{bmatrix}\right)$$

由正态分布的条件概率分布公式容易推出 $f(x^*)$ 的均值与方差函数如下：

$$\mathbf{m}_n(x^*) = \mathbf{m}(x^*) + \mathbf{k}(x^*)^T(\mathbf{K} + \sigma^2\mathbf{I})^{-1}(\mathbf{y} - \mathbf{m}(\mathbf{x})) \tag{4}$$

$$\sigma_n^2(x^*) = k(\mathbf{x}, \mathbf{x}) - \mathbf{k}(\mathbf{x})^T(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{k}(\mathbf{x}) \tag{5}$$

公式中的 $\mathbf{k}(x^*)$ 是 x^* 与 $\mathbf{x}_{1:n}$ 之间的协方差向量。值得一提的是，若我们对目标函数并没有十分具体的先验信息，则先验均值函数 $\mathbf{m}$ 一般直接取 $\mathbf{0}$。有了高斯过程模型后，我们对于搜索空间上的每一个点均能获得其预测均值和预测方差，这给采集函数的设计提供了重要基础。

2.2 采集函数

本文中的实验将采集函数聚焦于 Expected improvement(EI) 及其变体。EI 是贝叶斯优化中最常用的采集函数之一，其主要思想在于选择相比当前最优值提升期望最大的点作为下一个观测点。具体来说，我们先定义改进函数 I(improvement function):

$$I(\mathbf{x}, v, \theta) := (v - f_n^*)\mathbb{I}(v > f_n^*) \tag{6}$$

其中 $v \sim \mathcal{N}(m_n(\mathbf{x}), \sigma^2(\mathbf{x}))$, $\mathbb{I}$ 为示性函数，θ 为超参数集。f_n^* 为当前最优值，$m_n(\mathbf{x}), \sigma_n(\mathbf{x})$ 分别为高斯过程模型对输入值 $\mathbf{x}$ 所返回的预测均值和预测方差。Φ 为标准正态分布的累积分布函数 (CDF)。对 v 求期望即得到采集函数 EI：

$$u(\mathbf{x}; D_{1:n}) := \mathbb{E}[I(\mathbf{x}, v, \theta)] = (m_n(\mathbf{x}) - f_n^*)\Phi\Big(\frac{m_n(\mathbf{x}) - f_n^*}{\sigma(\mathbf{x})}\Big) + \sigma_n(\mathbf{x})\phi\Big(\frac{m_n(\mathbf{x}) - f_n^*}{\sigma_n(\mathbf{x})}\Big) \tag{7}$$

其中 Φ 和 ϕ 分别为标准正态分布的累积分布 (CDF) 和概率密度分布 (PDF)。当目标函数存在噪声时，当前最优值 f_n^* 并非真实值。在噪声较大时，f_n^* 的取值可能会较为极端，导致 EI 错误地认为搜索空间内的绝大部分点提升期望极小，从而使得 EI 过度倾向于探索而非开发。一个最直观的解决方式是用已观测点中代理模型的最小预测均值 m_n^* 来代替 f_n^*。这实际上是将 EI 从优化带噪声的最优观测值改进为优化代理模型的最优预测值，从而提升算法对噪声的鲁棒性。我们将改进后的 EI 称之为 EIm，其表达式如下：

$$u'(\mathbf{x}; D_{1:n}) := (m_n(\mathbf{x}) - m_n^*)\Phi\Big(\frac{m_n(\mathbf{x}) - m_n^*}{\sigma(\mathbf{x})}\Big) + \sigma_n(\mathbf{x})\phi\Big(\frac{m_n(\mathbf{x}) - m_n^*}{\sigma_n(\mathbf{x})}\Big) \tag{8}$$

μ^+ 为当前最优值，$\mu_n(\mathbf{x}), \sigma_n(\mathbf{x})$ 分别为高斯过程模型对输入值 $\mathbf{x}$ 所返回的预测均值和预测方差。Φ 为标准正态分布的累积分布函数 (CDF)。

In subsequent experiments, we will use EI for the output method based on observed values; and use EIm for the output method based on predicted values,.

3 测试函数及实验设置

3.1 BBOB 基准环境

本文的测试函数选自 BBOB 无噪声测试函数组的 24 个测试函数，在其基础上添加噪声作为我们的测试基准问题。我们将这些函数编号为 F1-F24，根据 BBOB 原文中的设置，这些函数被划分为 5 种类型：可分离函数 (F1-F5), 条件敏感度低或适中的函数 (F6-F9), 高条件敏感度单峰函数 (F10-F14), 有充分全局结构的多峰函数 (F15-F19), 全局结构较弱的的多峰函数 (F20-F24)。BBOB 测试函数组中的所有函数都可以通过平移或旋转等变化来获得新的测试函数。在本文的后续实验中，我们预先将这 24 个无噪声的函数固定下来，这些函数的最优值，标准差以及简单的描述见表 1。对于它们的更为详细的信息可以参考 [2]。

2D 函数	最优值	标准差	特点
F1	79.48	12.57	sphere function,unimodal,presumably the most easy continuous domain ssearch problem
F2	66.95	10044455.67	Globally quadratic and ill-conditioned(about 10^6) function with smooth local irregularities.Conditioning is about 10^6
F3	77.66	418.57	Highly multimodal function with a comparatively regular structure for the placement of the optima.
F4	77.66	171.86	Highly multimodal function with a structured but highly asymmetric placement of the optima.
F5	66.71	29.02	Purely linear function,solution is on the domain boundary
F6	65.87	237396.11	Unimodal,highly asymmetric function,
F7	92.94	431.55	unimodal, non-separable, conditioning is about 100.The function consists of many plateaus of different sizes.
F8	98.62	46480.79	(Rosenbrock function) in larger dimensions the function has a local optimum with an attraction volume of about 25%
F9	65.61	21715.05	rotated version of the previously defined f8.
F10	59.13	9634378.45	rotated version of the previously defined f2.
F11	76.27	21241083.28	A single direction in search space is a 1000 times more sensitive than all others.Conditioning is about 10^6
F12	56.61	9607260659	conditioning is about 10^6, rotated, unimodal
F13	68.42	449.99	Resembles f12 with a non-differentiable bottom of valley
F14	77.31	41.04	The sensitivies of the z_i-variables become more and more different when approaching the optimum
F15	70.03	521.74	Prototypical highly multimodal function which has originally a very regular and symmetric structure for the placement of the optima.
F16	71.35	79.07	Highly rugged and moderately repetitive landscape, where the global optimum is not unique.
F17	69.83	19.00	A highly multimodal function where frequency and amplitude of the modulation vary.Conditioning is low
F18	119.54	308.17	Moderately ill-conditioned counterpart to f17
F19	71.69	74.72	Resembling the Rosenbrock function in a highly multimodal way.
F20	71.29	45818.02	The most prominent 2D minima are located comparatively close to the corners of the unpenalized search area.
F21	124.08	12.21	The function consists of 101 optima with position and height being unrelated and randomly chosen.
F22	51.57	24.05	The function consists of 21 optima with position and height being unrelated and randomly chosen.Conditioning is about 1000
F23	85.39	20.12	Highly rugged and highly repetitive function with more than 10D global optima.
F24	93.30	18.18	Highly multimodal function with two funnels.

表 1: BBOB functions

3.2 实验设置

- **维度与预算：**实验中的函数维度分为 2 维与 4 维。对于二维问题，无噪声条件下单次实验总预算为 100 个观测点，有噪声时增加为 300 个。对于四维问题，无噪声条件下总预算为 200，带噪声条件下增加为 400。

- **初始化：**高斯过程建模的初始点数量我们统一设置为总预算的 10%，并采用拉丁超立方采样来在搜索区域内随机采点。

- **评估指标:**我们定义损失率 =(优化算法所得结果的真实值-真实最优值)/真实最优值 ×100% 来作为算法的优化性能评估指标。注意到我们测试的 24 个函数最小值均为正数且不趋近于 0，因此该损失率是良好定义的。类似地我们也可以定义两个算法之间的相对损失率 = (优化算法 1 所得结果真实值-优化算法 2 所得结果真实值)/优化算法 2 所得结果真实值。若该相对损失率为负值，则代表算法 1 结果优于算法 2.

- **随机性处理：**为了使结果更具一般性，我们每组实验均会取不同的随机种子重复 30 次后对结果取均值。贝叶斯优化过程中的随机性主要体现在初始点位置，采集函数内部优化两方面。

- **噪声设置：**本文实验中给测试函数添加的噪声均为高斯白噪声，其均值为 0，标准差设置为原函数标准差的 5%,20%,50% 作为不同程度的噪声大小 (小，中，大)。

- **采集函数及输出结果：**实验中会使用两种形式的采集函数：基于观测值优化的 EI 以及基于预测值优化的 EIm。如前文所述，贝叶斯优化是一种序贯优化，每一次迭代 EI 或者 EIm 都会给出一个新的观测点，这些点的集

合则会构成一个序列，我们分别称之为序列 EI 以及序列 EIm。同理，对于我们讨论的三种输出方式 (直接输出最小观测值，输出已观测点中预测均值最小观测点，输出全局预测均值最小点) 在每一次迭代同样会输出一个观测点，从而形成三个不同序列，我们分别简称为 $obs, obs\_M, total\_M$，每个序列的最后一个点即为算法在该次实验中给出的最终优化结果，分别称为 $output1, output2, output3$。

4 实验结果及分析

4.1 无噪声实验结果

我们首先关注无噪声的实验结果。在无噪声的情况下，由于观测点的观测值均为真实值，故序列 obs, obs_M 两者完全相同。同理，$output1, output2$ 也完全一致。我们只需对比序列 $obs, total\_M$ 来分析预测性的序列是否能够包含较好的点，以及 $output2$ 是否能够将其中较好的点作为最终结果输出出来。当然，从常理上判断 $output1$ 在无噪声的条件下是最合理的输出方式，我们的无噪声实验主要是想探究无噪声条件下如果使用预测性的输出方式其优化效果与观测性的输出方式之间有多大的差距。具体的实验结果见表 2:

表格中的前三列分别是序列 $obs, total$ 以及输出 $output3$ 相对于各个函数全局最优值的损失率。后两列则分别是序列 $total$ 与输出 $output3$ 与序列 obs 的相对损失率。从第四列可以很直观地看出，对于大多数测试函数，序列 obs 与序列 $total$ 中的最优值之间差距不大 (基本都在 1% 以内)。这说明对于大多数的测试函数来说，基于观测值或者预测值的输出方式都能够找到不错的点。同时，我们也可以看到基于预测性的 $total$ 序列在函数 F2, F10,F11,F12 上的优化表现十分糟糕。回顾表 1 可以发现，这些函数都有一个共同的特点：条件敏感度特别高 (均超过 10^6)。这些函数对应的标准差也具有很大的数值，这使得对应的噪声数值也较大。过大的噪声很容易覆盖掉函数原本的结构特征，使得高斯过程的建模变得异常艰难，无法起到很好的预测作用。事实上，观察表 2 我们可以发现，这四个函数对应的序列 obs 的损失率也相对较大，这说明普通的贝叶斯优化在面对这类问题时同样也会受到建模精度的影响。因此我们可以得到结论：对于无噪声问题，当高斯过程模型能够对目标函数进行良好建模时，基于观测性的输出方式与预测性的输出方式都能够找到较好的点，但若高斯过程的建模精度不高，则使用预测性的输出方式是不可接受的。虽然对于大多数函数来说预测性的输出方式能够找到不错的点，但从第五列的结果来看，预测性的方式输出的最终结果并不能保证将找到的最优的点给输出出来。$output3$ 相对 obs 的损失率大多在 10% 以上，对于更为复杂的函数 F10，F12 等，其优化结果更是难以接受。因此，对于无噪声的问题，我们确实应该直接使用基于观测值的贝叶斯优化结果输出方式。(所有测试函数中 $output3$ 相对于 obs 的损失率均大于 0)

2D 函数	obs	total	optput3	total_vs_obs	output3_vs_obs
F1	0.00%	0.00%	0.00%	0.00%	0.00%
F2	5.71%	45.31%	59.70%	36.90%	50.71%
F3	10.22%	8.40%	27.81%	-1.32%	16.80%
F4	5.01%	4.08%	18.19%	-0.85%	12.62%
F5	0.00%	0.00%	0.00%	0.00%	0.00%
F6	1.52%	2.58%	3.45%	1.05%	1.89%
F7	0.08%	0.08%	0.18%	0.00%	0.10%
F8	0.07%	0.20%	1.75%	0.13%	1.68%
F9	0.12%	0.69%	3.66%	0.57%	3.54%
F10	15.30%	41.30%	1282.36%	25.07%	1133.40%
F11	8.50%	48.75%	719.52%	38.60%	656.81%
F12	1645.98%	21840.61%	2907234.96%	2905.92%	1209546.08%
F13	0.96%	0.56%	1.78%	-0.38%	0.81%
F14	0.01%	0.01%	0.02%	0.00%	0.02%
F15	5.26%	5.60%	23.30%	0.36%	17.33%
F16	0.74%	0.60%	26.55%	-0.14%	25.71%
F17	0.52%	0.38%	4.43%	-0.14%	3.89%
F18	0.76%	0.60%	11.46%	-0.16%	10.62%
F19	0.25%	0.30%	10.95%	0.05%	10.68%
F20	1.64%	2.41%	5.11%	0.76%	3.41%
F21	0.07%	0.05%	0.14%	-0.02%	0.07%
F22	0.46%	0.40%	0.75%	-0.06%	0.29%
F23	4.48%	4.53%	50.64%	0.10%	44.24%
F24	5.11%	4.56%	27.33%	-0.49%	21.20%

表 2: 2D noiseless results

参考文献

[1] Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P. Adams, and Nando de Freitas. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2016.

[2] Nikolaus Hansen, Anne Auger, Steffen Finck, and Raymond Ros. Real-parameter black-box optimization benchmarking bbob-2010 : Experimental setup. 2010.