

An interpretation of Bayesian global optimization as proposed by Snyman and Fatti [1]

Dirk Munro (Hamburg, Germany)

January 3, 2022

1 Can we develop confidence in the global optimality of a solution?

Assume that we have an optimization problem and a search algorithm (e.g. gradient-based) to solve it. That is, for a given starting position \mathbf{x}_0 the algorithm will return a stationary point \mathbf{x}_* , a candidate optimum from all the potential optima—*i.e.* local optima, expect for the one global optima.

Assume that the optimization problem has a finite but unknown number of optima \mathbf{x}_*^j , with $j = 1, 2, \dots, s$, and that from a particular starting point \mathbf{x}_0 , the algorithm will converge to one of them. This procedure which transforms a given \mathbf{x}_0 to one of the the optima \mathbf{x}_*^j is deterministic, yet unpredictable. In other words, given a new \mathbf{x}_0 (pulled ‘from a hat’, so to speak), we do not know *a priori* which \mathbf{x}_*^j will be returned; however, given the same \mathbf{x}_0 again, the same \mathbf{x}_*^j (from before) will be returned.

Lets say an an ‘experiment’ consists of sampling a random starting point \mathbf{x}_0 from the design domain, and running the optimization algorithm. There is hence a probability P_*^j that a particular \mathbf{x}_*^j will be returned. Also, there is a probability P_* that the (global) optimum will be returned.

What is the probability P_* of finding the global optimum from a random starting point? If we had an estimate of the total number of optima s , then a reasonable estimate may be $1/s$...? Or rather, the flexible form of the Beta distribution may be assumed

$$\mathbf{p}[P_*] = \frac{1}{\beta[a, b]} P_*^{a-1} (1 - P_*)^{b-1}, \quad (1)$$

which reflects the *probability density function* \mathbf{p} of the probability P_* of finding the global optimum from a random starting position—see for example Figure 1. An expected value of P_* given an a and b , and other statistical measures, is implied by (1).

Bayes’ theorem permits us to update the prior probability density function $\mathbf{p}[P_*]$, given the results from a number of experiments. Assume (for now) that we know the number of ‘successes’ r , after a total number of experiments n . That is, running the algorithm from n random starting positions, we assume that we can measure the number of times r the global optimum was found (r ‘successes’)¹. Given n and r , the prior probability

¹The severity/subtlety of this assumption—*i.e.*, knowing which is the global optimum, and hence being able to measure the amount of times it is found—is addressed later.

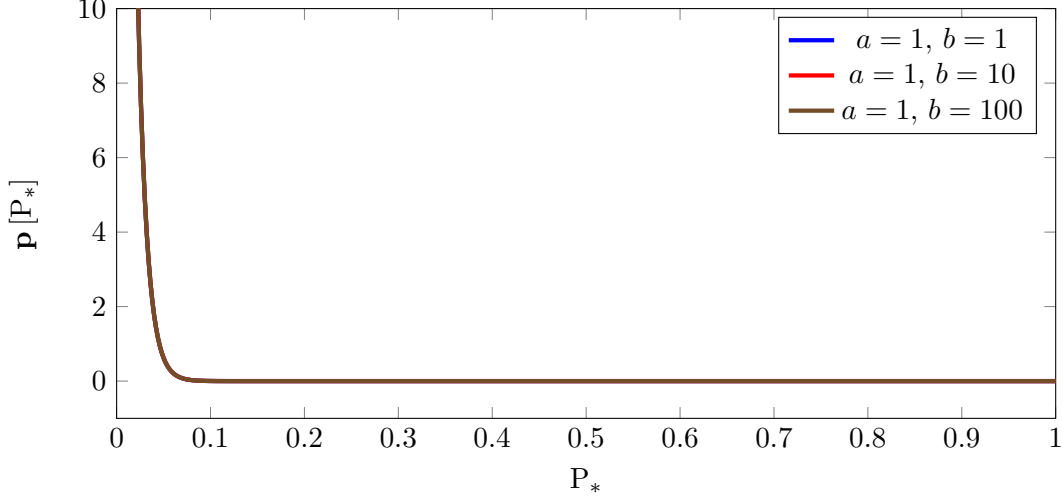


Figure 1: Example Beta distributions (probability density functions) of P_* , for representative values of a and b .

density function of P_* may be updated as per Bayes

$$\mathbf{p}[P_* | [n, r]] = \mathbf{p}[r | [n, P_*]] \mathbf{p}[P_*] \Big/ \int_0^1 \mathbf{p}[r | [n, P_*]] \mathbf{p}[P_*] dP_* \quad (2)$$

The new term $\mathbf{p}[r | [n, P_*]]$ is the probability density function of the number of ‘successes’ r , given n experiments and a prior probability P_* . This probability density function is thus of the binomial form

$$\mathbf{p}[r | [P_*, n]] = \frac{n!}{r!(n-r)!} P_*^r (1 - P_*)^{n-r}. \quad (3)$$

In the end, we want to have a measure of confidence in whether we may have found the (global) optimum of the optimization problem, given n experiments, r ‘successes’, and the prior P_* . On the one hand, this can be expressed as the probability of having found the true optimum *at least once*, given n and P_*

$$P[r_* \geq 1 | [n, P_*]] = 1 - (1 - P_*)^n, \quad (4)$$

which follows from the binomial distribution (3) of the complementary probability: 1 minus the probability of not finding the true optimum at all ($r = 0$, after n experiments). On the other hand, we have a probability density function $\mathbf{p}[P_* | [n, r]]$, updated with the results r of n experiments, as per Bayes’ theorem (2). Applying a change-of-variable [2] to $\mathbf{p}[P_* | \dots]$ from Eq. (2), with Eq. (4) the transformation function, the probability of having found the optimum at least once, in terms of n and r , is

$$P[r_* \geq 1 | [n, r]] = \int_0^1 P[r_* \geq 1 | [n, P_*]] \mathbf{p}[P_* | [n, r]] dP_*. \quad (5)$$

Now, the issue of being able to count the number of ‘successes’ r_* from n experiments is addressed. The objective value of a particular optima may be evaluated, $F_j [\mathbf{x}_*^j]$, and values amongst different optima may be compared: *e.g.* is $F_c [\mathbf{x}_*^c]$ a better optima than $F_d [\mathbf{x}_*^d]$? Hence, from n experiments, we can determine what was relatively the best optimum solution found χ_* , and we can count how many solutions ρ_* (‘relative successes’) correspond to this χ_* . If we assume that the probability of finding the global optimum from a random starting point P_* , is *at least* the probability of finding any other P_*^j

$$P_* \geq P_*^j \forall j, \quad (6)$$

then it follows that

$$P[r_* \geq 1 | [n, r]] \geq P[\rho_* \geq 1 | [n, r]] . \quad (7)$$

That is, the probability of having had at least one absolute ‘success’ r_* , after n experiments, is greater than the probability of having at least one ‘relative success’ ρ_* . This is property that may have to be justified in the optimization algorithm; *e.g.*, move-limits should be sufficiently large?

Substitution of (4) and (2) in (5), and, in turn, substitution of the binomial (3) and beta (1) distributions, permits straight-forward cancellations and simplification [1] to

$$P_f = P[r_* \geq 1 | [n, r]] \geq P[\rho_* \geq 1 | [n, r]] = \\ 1 - \beta[r + a, 2n - r + b] / \beta[r + a, n - r + b] = \frac{\Gamma[n + a + b] \Gamma[2n - r + b]}{\Gamma[2n + a + b] \Gamma[n - r + b]} . \quad (8)$$

Further cancellations are made in the factorials, and, in numerical terms, the calculation is dealt with as per the Python function given verbatim below.

In Table 1 some example calculations are given with a severely pessimistic estimate of $\mathbf{p}[P_*]$, with $a = 1$ and $b = 1000$ (see Fig. 1). We can see how the confidence in having found the global optimum P_f changes for different values of n and r . For example, with 10 experiments (n), even if we have a success rater of 1 in 2 ($r = 5$), our confidence barely rounds-up to 1 %. However, if we do 100 experiments, and we have only between 5 and 10 successes, then the confidence of having found the global optimum increases to about 50%. This is not bad, given a prior expectation of $\frac{1}{1+1000} \approx \frac{1}{1000}$ (0.1 %). Given 200 experiments, and 10 successes, our confidence increases to a fairly comfortable 80%, *etc.*.

```
#
def P_f(ni,ri,a,b):
#
#   see Bolton (2004)
#
    tmp=1e0; abar=a+b-1; bbar=b-ri-1
    for i in range(1,ni+1):
```

Table 1: Example calculations using Eq. (8), with $a = 1$ and $b = 1000$ in the prior $\mathbf{p}[P_*]$

n	r	P_f	n	r	P_f	n	r	P_f	n	r	P_f
10	1	0.02	100	1	0.16	200	1	0.27	1000	1	0.56
10	2	0.03	100	2	0.23	200	2	0.37	1000	2	0.70
10	3	0.04	100	3	0.30	200	3	0.46	1000	3	0.80
10	4	0.05	100	4	0.35	200	4	0.54	1000	4	0.87
10	5	0.06	100	5	0.41	200	5	0.60	1000	5	0.91
10	10	0.10	100	10	0.62	200	10	0.82	1000	10	0.99

```

        tmp=tmp*(ni+i+bbar)/(ni+i+abar)
#
    return 1e0-tmp
#

```

2 Topology optimization experiments

References

- [1] J.A. Snyman and L.P. Fatti. A multi-start global minimization algorithm with dynamic search trajectories. *Journal of optimization theory and applications*, 54(1):121–141, 1987.
- [2] F.M. Dekking, C. Kraaikamp, H. P. Lopuhaä, and L. E. Meester. *A modern introduction to probability and statistics: understanding why and how*. Springer Science & Business Media, 2005.