What do you Mean?

The Role of the Mean Function in Bayesian Optimisation

Supplementary Material

GEORGE DE ATH, University of Exeter, United Kingdom

JONATHAN E. FIELDSEND, University of Exeter, United Kingdom

RICHARD M. EVERSON, University of Exeter, United Kingdom

ACM Reference Format:

George De Ath, Jonathan E. Fieldsend, and Richard M. Everson. 2020. What do you Mean? The Role of the Mean Function in Bayesian Optimisation: Supplementary Material. In *Genetic and Evolutionary Computation Conference Companion (GECCO '20 Companion)*, July 8–12, 2020, Cancún, Mexico. ACM, New York, NY, USA, 4 pages. <https://doi.org/10.1145/3377929.3398118>

A ADDITIONAL RESULTS

In this section we show convergence plots for the Branin and GoldsteinPrice synthetic test functions for both the EI and UCB acquisition functions, as well as the results tables for UCB on the synthetic functions, and EI and UCB on the three real-world problems.

Fig. 1. Convergence plots for the Branin and GoldsteinPrice test problem using EI (left) and UCB (right). Each plot shows the median regret, with shading representing the interquartile range across the 51 runs.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

GECCO '20 Companion, July 8–12, 2020, Cancún, Mexico

© 2020 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-7127-8/20/07...\$15.00

<https://doi.org/10.1145/3377929.3398118>

Table 1. Mean function performance using the UCB acquisition function. Median regret (*column left*) and median absolute deviation from the median (MAD, *column right*) after 200 function evaluations across the 51 runs. The method with the lowest median performance is shown in dark grey, with those statistically equivalent shown in light grey.

Mean function	Branin (2)		Eggholder (2)		GoldsteinPrice (2)		SixHumpCamel (2)		Shekel (4)	
	Median	MAD	Median	MAD	Median	MAD	Median	MAD	Median	MAD
Arithmetic	1.27×10^{-4}	1.47×10^{-4}	2.47	2.77	1.37	1.24	7.94×10^{-5}	7.12×10^{-5}	4.20×10^{-2}	3.38×10^{-2}
Median	1.11×10^{-4}	1.44×10^{-4}	3.21	2.76	1.21	9.88×10^{-1}	7.44×10^{-5}	7.44×10^{-5}	2.11×10^{-4}	2.89×10^{-4}
Min	1.27×10^{-4}	1.57×10^{-4}	4.09	3.09	1.12	9.27×10^{-1}	8.57×10^{-5}	9.33×10^{-5}	7.88	6.60×10^{-1}
Max	1.46×10^{-4}	1.55×10^{-4}	3.31	3.01	1.04	1.09	8.75×10^{-5}	7.29×10^{-5}	3.63×10^{-4}	4.03×10^{-4}
Linear	9.20×10^{-5}	8.31×10^{-5}	2.66	2.34	1.36	8.95×10^{-1}	4.53×10^{-4}	5.47×10^{-4}	6.47	7.89×10^{-2}
Quadratic	3.05×10^{-3}	4.36×10^{-3}	3.08	4.15	1.78	1.59	2.15×10^{-2}	2.01×10^{-2}	4.98×10^{-1}	7.31×10^{-1}
RandomForest	4.30×10^{-3}	4.55×10^{-3}	2.96	3.30	2.18	1.49	1.16×10^{-3}	1.40×10^{-3}	6.61	1.76
RBF	1.59×10^{-3}	2.21×10^{-3}	5.05	4.05	5.36	6.16	9.78×10^{-3}	1.43×10^{-2}	6.64	1.60

Mean function	Ackley (5)		Hartmann6 (6)		Michalewicz (10)		Rosenbrock (10)		StyblinskiTang (10)	
	Median	MAD	Median	MAD	Median	MAD	Median	MAD	Median	MAD
Arithmetic	3.53	6.53×10^{-1}	9.93×10^{-2}	1.22×10^{-1}	7.90×10^{-2}	9.42×10^{-2}	2.95×10^4	2.36×10^4	1.18×10^2	2.40×10^1
Median	2.37	5.60×10^{-1}	1.94×10^{-2}	2.70×10^{-2}	4.84×10^{-2}	5.89×10^{-2}	2.00×10^4	1.61×10^4	1.71×10^2	2.97×10^1
Min	7.06	1.49	3.03×10^{-1}	1.29×10^{-1}	1.74	6.51×10^{-1}	2.45×10^4	2.21×10^4	1.98×10^2	2.57×10^1
Max	2.21	4.51×10^{-1}	4.40×10^{-2}	6.48×10^{-2}	3.92×10^{-1}	3.59×10^{-1}	8.22×10^3	8.89×10^3	1.91×10^2	2.52×10^1
Linear	3.35	8.75×10^{-1}	6.31×10^{-2}	8.03×10^{-2}	9.27×10^{-1}	1.13	2.20×10^4	1.20×10^4	1.77×10^2	2.94×10^1
Quadratic	4.75	1.71	1.87×10^{-1}	1.84×10^{-1}	6.54×10^{-1}	5.72×10^{-1}	2.06×10^3	1.10×10^3	1.06×10^2	2.47×10^1
RandomForest	4.42	5.62×10^{-1}	8.03×10^{-2}	6.30×10^{-2}	1.51×10^{-1}	1.59×10^{-1}	8.27×10^3	4.96×10^3	7.21×10^1	1.75×10^1
RBF	3.67	1.05	1.33×10^{-1}	6.83×10^{-2}	7.21×10^{-1}	5.60×10^{-1}	2.87×10^3	1.18×10^3	1.18×10^2	2.12×10^1

Table 2. Mean function performance using the EI (*left table*) and UCB (*right table*) acquisition functions. Median regret (*column left*) and median absolute deviation from the median (MAD, *column right*) after 200 function evaluations across the 51 runs are shown in paired columns. The method with the lowest median performance is shown in dark grey, with those statistically equivalent in light grey.

Mean function	PUSH4 (4)		PUSH8 (8)	
	Median	MAD	Median	MAD
Arithmetic	1.53×10^{-1}	1.16×10^{-1}	2.77	1.93
Median	1.54×10^{-1}	6.53×10^{-2}	2.19	1.58
Min	2.39×10^{-1}	1.60×10^{-1}	3.53	1.44
Max	1.28×10^{-1}	9.75×10^{-2}	2.66	1.83
Linear	2.62×10^{-1}	2.10×10^{-1}	3.39	2.05
Quadratic	1.92×10^{-1}	1.28×10^{-1}	3.16	1.83
RandomForest	6.56×10^{-2}	5.71×10^{-2}	3.07	2.24
RBF	1.64×10^{-1}	1.19×10^{-1}	3.35	2.29

Mean function	PUSH4 (4)		PUSH8 (8)	
	Median	MAD	Median	MAD
Arithmetic	3.42×10^{-1}	2.18×10^{-1}	3.95	1.57
Median	5.01×10^{-1}	3.06×10^{-1}	3.64	2.05
Min	4.36×10^{-1}	3.69×10^{-1}	4.37	1.72
Max	4.25×10^{-1}	3.20×10^{-1}	4.57	1.74
Linear	7.48×10^{-1}	5.47×10^{-1}	4.00	2.70
Quadratic	4.69×10^{-1}	4.37×10^{-1}	3.79	1.94
RandomForest	1.05×10^{-1}	6.92×10^{-2}	2.36	2.11
RBF	1.39×10^{-1}	9.29×10^{-2}	3.66	2.43

Table 3. Mean function performance using the EI (*left table*) and UCB (*right table*) acquisition functions. Median regret (*column left*) and median absolute deviation from the median (MAD, *column right*) after 200 function evaluations across the 51 runs are shown in paired columns. The method with the lowest median performance is shown in dark grey, with those statistically equivalent in light grey.

Mean function	PitzDaily (10)	
	Median	MAD
Arithmetic	8.39×10^{-2}	1.12×10^{-3}
Median	8.60×10^{-2}	3.23×10^{-3}
Min	8.43×10^{-2}	1.39×10^{-3}
Max	8.38×10^{-2}	1.38×10^{-3}
Linear	8.50×10^{-2}	2.24×10^{-3}
Quadratic	8.57×10^{-2}	3.07×10^{-3}
RandomForest	8.84×10^{-2}	6.38×10^{-3}
RBF	8.79×10^{-2}	6.23×10^{-3}

Mean function	PitzDaily (10)	
	Median	MAD
Arithmetic	8.41×10^{-2}	3.92×10^{-4}
Median	8.40×10^{-2}	1.96×10^{-4}
Min	8.39×10^{-2}	2.01×10^{-4}
Max	8.40×10^{-2}	3.07×10^{-4}
Linear	8.41×10^{-2}	3.98×10^{-4}
Quadratic	8.43×10^{-2}	9.23×10^{-4}
RandomForest	8.77×10^{-2}	4.69×10^{-3}
RBF	8.96×10^{-2}	8.23×10^{-3}

B GP MODELLING ERROR USING DIFFERENT MEAN FUNCTIONS

In this section we present a short investigation into the modelling capability of the Gaussian process (GP) when combined with different mean functions. In order to quantify the modelling capability of a GP using a prior mean function on a test function $f : \mathbb{R}^d \mapsto \mathbb{R}$, we created GP models of f using training data taken from the optimisation runs, evaluated a set of $N = 1000$ test locations, and calculated the model’s normalised root mean squared error (NRMSE).

More precisely, for a given test problem f , a GP model using the mean function to be evaluated was created using the first 100 locations (including initial LHS samples) for each of the mean function’s 51 optimisation runs using the EI acquisition function. A set of $N = 1000$ test locations $X = \{\mathbf{x}_n\}_{n=1}^N$ were chosen via Latin hypercube sampling and the model’s prediction NRMSE between the true function values f_n and predicted values from the surrogate model $\hat{f}_n$ were calculated for each of the 51 models. The NRMSE is defined as

$$\text{NRMSE} = \frac{\sqrt{\frac{1}{N} \sum_{n=1}^N (f_n - \hat{f}_n)^2}}{f_{\max} - f_{\min}}, \quad (1)$$

where $f_{\max}$ and $f_{\min}$ are the largest and smallest values of the test locations evaluated using the real function. The test locations X were paired across models corresponding to optimisation runs using the same training data so that a statistical comparison could be undertaken.

Table 4 shows the median NRMSE over the 51 GP models created using each mean function for the synthetic and robot pushing test problems. The method with the lowest median NRMSE on each function is highlighted in dark grey, and those highlighted in light grey are statistically equivalent to the best method according to a one-sided paired Wilcoxon signed-rank test with Holm-Bonferroni correction ($p \geq 0.05$). The RBF mean function provides the best (or equivalent to the best) modelling error for the evaluated test locations on 10 of the 12 evaluated functions. Interestingly, this does not correspond the same performance in Bayesian optimisation (recall Table 1 in the main paper). We suspect this is because the experiments are evaluating modelling of the entire function, whereas BO only requires the GP model to reflect the true function within the region that the minima reside.

Further work in quantifying the modelling error in BO could be carried out by only calculating the NRMSE within a small hypercube surrounding the global optimum. However, the size of the hypercube required to accurately calculate the modelling error of the important features for BO may vary between problems and therefore additional evaluations may be required of sets of test data sampled from hypercubes of different sizes.

	Arith	Med	Min	Max	Lin	Quad	RF	RBF
Branin	0.271	0.271	0.271	0.263	0.270	0.263	0.121	0.114
Eggholder	0.217	0.165	0.216	0.177	0.161	0.163	0.198	0.166
GoldsteinPrice	0.136	0.136	0.136	0.136	0.136	0.136	0.069	0.136
SixHumpCamel	0.232	0.231	0.231	0.223	0.215	0.216	0.142	0.228
Shekel	0.265	0.215	0.494	0.058	0.083	0.094	0.142	0.053
Ackley	0.663	0.716	0.316	0.899	0.107	0.378	0.334	0.117
Hartmann6	0.222	0.394	0.184	0.154	0.140	0.185	0.195	0.097
Michalewicz	0.271	0.311	0.417	0.316	0.205	0.193	0.360	0.190
Rosenbrock	0.362	0.368	0.371	0.270	0.257	0.151	0.355	0.157
StyblinskiTang	0.325	0.347	0.360	0.151	0.267	0.199	0.220	0.168
PUSH4	0.202	0.200	0.256	0.254	0.153	0.152	0.222	0.151
PUSH8	0.205	0.202	0.238	0.168	0.136	0.129	0.191	0.125

Table 4. Median NRMSE modelling prediction error taken over across 51 GP models using the the 8 evaluated mean functions on the ten synthetic and two robot pushing functions. The models with the lowest NRMSE are shown in dark grey, with those statistically equivalent to them shown in light grey.