ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

Joint Entropy Search for Multi-objective Bayesian Optimization with Constraints and Multiple Fidelities

Daniel Fernández-Sánchez[1] and Daniel Hernández-Lobato[1] †

1- Universidad Autónoma de Madrid - Computer Science Department
Francisco Tomás y Valiente 11, 28049, Madrid - Spain

Abstract. Bayesian optimization (BO) methods solve problems with several black-box objectives and constraints. Each black-box is expensive to evaluate and lacks a closed-form. They use a model of each black-box to guide the search for the problem's solution. Sometimes, however, the black-boxes may be evaluated at different fidelity levels. A lower fidelity is simply a cheap proxy of the corresponding black-box. Thus, lower fidelities that correlate with the actual black-box can be used to reduce the optimization cost. We propose Joint Entropy Search for Multi-Fidelity and Multi-objective Bayesian Optimization with Constraints (MF-JESMOC), a BO method for solving the aforementioned problems. It chooses the next point and fidelity level at which to evaluate the black-boxes as the one that is expected to reduce the most the joint entropy of the Pareto set and the Pareto front, normalized by the fidelity's cost. Deep Gaussian processes are used to model each black-box and dependencies between fidelities. In our experiments, MF-JESMOC outperforms other state-of-the-art methods for multi-objective BO with constraints and different fidelity levels.

1 Introduction

We aim at minimizing K objectives $f_1(\mathbf{x}), \ldots, f_K(\mathbf{x})$, under J constraints $c_1(\mathbf{x}) \geq 0, \ldots, c_J(\mathbf{x}) \geq 0$. We assume $\mathbf{x} \in \mathcal{X}$ and $\mathcal{X} \subset \mathbb{R}^D$. The objectives are expected to be conflictive and there is not a single optimum, but a set of optimal trade-offs among the objectives that meet the constraints. This set is called the Pareto set $\mathcal{X}^\star$ and evaluating at this set the objectives yields the Pareto front $\mathcal{Y}^\star$. The Pareto set $\mathcal{X}^\star$ contains feasible non-dominated points, *i.e.*, $\mathcal{X}^\star \subset \mathcal{F}$, with $\mathcal{F} = \{\mathbf{x} \in \mathcal{X} : c_j(\mathbf{x}) \geq 0, \forall j\}$. We say that $\mathbf{x}$ dominates $\mathbf{x}'$ if $f_k(\mathbf{x}) \leq f_k(\mathbf{x}') \, \forall k$ with at least one inequality being strict. Specifically, we want to solve:

$$\min_{\mathbf{x} \in \mathcal{X}} \quad f_1(\mathbf{x}), \ldots, f_K(\mathbf{x}) \quad s.t. \quad c_1(\mathbf{x}) \geq 0, \ldots, c_J(\mathbf{x}) \geq 0 \,. \tag{1}$$

We also assume that the objectives and the constraints are black-boxes. That is, they lack a closed-form expression and their evaluations are potentially noisy and very expensive to compute. *E.g.*, we may be interested in simultaneously

†The authors acknowledge financial support from projects PID2019-106827GB-I00 and PID2022-139856NB-I00, funded by MCIN and from the Autonomous Community of Madrid (ELLIS Unit Madrid). They also acknowledge the use of the facilities of Centro de Computación Científica, UAM.

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

minimizing the error of a deep neural network (DNN) and its prediction time, so that once codified on a chip, its power consumption is below some level. Evaluating the black-boxes in this case involves training the DNN on some data, measuring its prediction time, and running a computer simulation, respectively.

Bayesian optimization (BO) methods solve these problems efficiently [1, 2]. They use a model of each black-box to guide the search for the problem's solution. Specifically, the model's predictive distribution identifies the potential values of each black-box in un-observed regions of $\mathcal{X}$. This uncertainty is then transformed into an acquisition function, whose maximum gives the next evaluation of the black-boxes. This process of data collection and model fitting repeats until a computational budget is met. After this, the model's predictive means are optimized to suggest the problem's solution. The key for BO success is that fitting the models and optimizing the acquisition function is very cheap compared to evaluating the black-boxes. In summary, BO methods carefully choose where to evaluate next, and often require fewer evaluations than other methods.

Sometimes it is possible to perform cheaper evaluations of the black-boxes at a lower fidelity level. *E.g.*, in the previous example, the DNN may be trained for only a few epochs. The error of that DNN will be larger, but hopefully similar to that of the fully trained DNN. Thus, lower fidelity evaluations are much cheaper to obtain, and are expected to be correlated with the actual black-box value. Using them in the BO process can reduce the optimization cost. Importantly, however, the goal is still to get $\mathcal{X}^\star$ and $\mathcal{Y}^\star$, corresponding to the highest fidelity.

Recent works have addressed multi-objective BO with constraints [1, 2] and multi-objective BO with multiple fidelities [3, 4, 5, 6]. However, multi-objective BO with constraints and multiple fidelities has received less attention, with the exception of [7]. Here, we propose Joint Entropy Search (JES) for Constrained Multi-objective Bayesian Optimization (CMOBO) with Multiple Fidelities. This method selects at each iteration the point and fidelity level that reduces the most the entropy of $\{\mathcal{X}^\star, \mathcal{Y}^\star\}$.

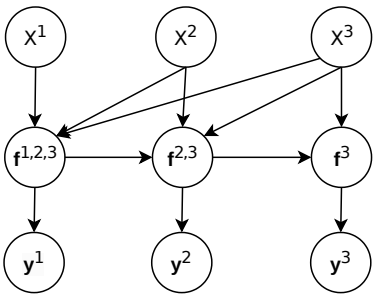

Fig. 1: MF-DGP for training.

2 JES for CMOBO with Multiple Fidelities

The key of any BO method is the model used to estimate potential black-box function values. We describe first our model and then our acquisition function.

2.1 Deep GPs for Multi-Fidelity Bayesian Optimization

Let $\mathcal{D} = \{(\mathbf{X}^1, \mathbf{y}^1), ..., (\mathbf{X}^t, \mathbf{y}^t), ..., (\mathbf{X}^T, \mathbf{y}^T)\}$ be a dataset of observations made at T different fidelities, with T the target fidelity and $t < T$ lower fidelity levels. We consider the multi-fidelity deep Gaussian Process (MF-DGP) of [8] to model

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

these data. MF-DGP uses a GP to model each fidelity. Let f^t be the latent function of fidelity t. The observation y_i^t at input $\mathbf{x}_i^t$ depends on $\mathbf{x}_i^t$, but also on f_i^{t-1}, *i.e.*, the previous fidelity level at $\mathbf{x}_i^t$. More precisely, it is assumed that $y_i^t = f^t(\mathbf{x}_i^t, f^{t-1}(\mathbf{x}_i^t, f^{t-2}(\mathbf{x}_i^t, \ldots))) + \epsilon_i^t$, with $\epsilon_i^t \sim \mathcal{N}(0, \sigma_t^2)$ and $y_i^1 = f^1(\mathbf{x}_i^1) + \epsilon_i^1$. This recursive evaluation introduces dependencies between the fidelities. Figure 1 shows the architecture of a MF-DGP, where the first layer is a regular GP.

A deep model with GPs leads to intractabilities [9]. Thus, MF-DGP uses a sparse variational approximation of a GP in each layer [10], and the parameters of the model are learned by maximizing the Evidence Lower Bound (ELBO):

$$\mathcal{L}_{\text{MF-DGP}} = \sum_{t=1}^{T} \sum_{i=1}^{N^t} \mathbb{E}_{q_t(f_i^t)}[\log p(y_i^t | f_i^t)] - \sum_{t=1}^{T} \mathbb{KL}(q_t(\mathbf{u}^t) || p(\mathbf{u}^t | \mathbf{Z}^t)) \,, \quad (2)$$

where f_i^t is the i-th value observed at fidelity t without noise, y_i^t is that value with noise, $\mathbf{u}^t$ are the process values at the inducing points $\mathbf{Z}^t$ at layer t, q_t is a variational distribution and $\mathbb{KL}(\cdot||\cdot)$ is the Kullback-Leibler divergence.

To model linear and non-linear dependencies between fidelities MF-DGP uses the following kernel at fidelity (layer) t:

$$\begin{aligned} k_t(\mathbf{x}_i^t, \mathbf{x}_j^t) = k_t^\rho(\mathbf{x}_i^t, \mathbf{x}_j^t; \theta_t^\rho) \Big[\hat{\sigma}_t^2 (f^{t-1}\left(\mathbf{x}_i^t\right) - c)^\intercal (f^{t-1}\left(\mathbf{x}_j^t\right) - c) \\ + k_t^{f-1}\left(f^{t-1}\left(\mathbf{x}_i^t\right), f^{t-1}\left(\mathbf{x}_j^t\right); \theta_t^{f-1}\right) \Big] + k_t^\delta(\mathbf{x}_i^t, \mathbf{x}_j^t; \theta_t^\delta) \,, \end{aligned} \quad (3)$$

where all $k(\cdot,\cdot)$ are RBF kernels and θ their parameters. k_t^ρ acts as an input-dependent scaling factor, k_t^{f-1} models non-linear dependencies w.r.t. the previous fidelity and $\hat{\sigma}_t^2$ and c are parameters used to capture linear dependencies. Finally, $k_t^\delta(\cdot,\cdot)$ is a bias term used to account for differences among fidelities.

2.2 MF-JESMOC's Acquisition Function

Joint entropy search (JES) is an acquisition function that chooses the next point as the one that reduces the entropy of the problem's solution $\{\mathcal{X}^\star, \mathcal{Y}^\star\}$ the most [11]. A reformulation of JES for our particular setting is:

$$\alpha(\mathbf{x}, t) = \left(H\left[p(\mathbf{y}^t | \mathcal{D}, \mathbf{x}^t)\right] - \mathbb{E}_{\mathcal{X}^\star, \mathcal{Y}^\star}\left[H\left[p(\mathbf{y}^t | \mathcal{D}, \mathbf{x}^t, \mathcal{X}^\star, \mathcal{Y}^\star)\right]\right] \right) C_t^{-1} \,, \quad (4)$$

where C_t is the cost of fidelity t, $\mathcal{D}$ is the collected data, $p(\mathbf{y}^t | \mathcal{D}, \mathbf{x}^t)$ is the predictive distribution of the black-boxes for fidelity t at $\mathbf{x}$, and $p(\mathbf{y}^t | \mathcal{D}, \mathbf{x}^t, \mathcal{X}^\star, \mathcal{Y}^\star)$ is the predictive distribution conditioned on the problem's solution. Finally, $H[\cdot]$ indicates entropy. Eq. (4) is the mutual information between $\{\mathcal{X}^\star, \mathcal{Y}^\star\}$ and $\mathbf{y}^t$ per cost. Thus, it favors choosing $\mathbf{x}$ and t where $\mathbf{y}^t$ is most informative about $\{\mathcal{X}^\star, \mathcal{Y}^\star\}$, per fidelity evaluation cost.

$H\left[p(\mathbf{y}^t | \mathcal{D}, \mathbf{x}^t)\right]$ in (4) is the entropy of the predictive distribution of each MF-DGP for the black-boxes. This entropy can be approximated by the entropy of a Gaussian distribution with the same moments. The moments can be obtained by propagating samples through each MF-DGP as in [9]. Moreover, since we assume independence among black-boxes, we only have to sum individual entropies.

The expectation in Eq. (4) can be approximated via Monte Carlo. We only have to generate samples of the highest fidelity of each black-box and optimize

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

them. The samples are obtained using a random Fourier features approximation of the sparse GPs in the MF-DGP, as in [12]. The samples are optimized using a grid of points as in [1]. In practice, we only generate a single sample of $\{\mathcal{X}^\star, \mathcal{Y}^\star\}$.

We also need to approximate the entropy of the conditional distribution, $p(\mathbf{y}^t|\mathcal{D}, \mathbf{x}^t, \mathcal{X}^\star, \mathcal{Y}^\star)$, as it is intractable. The conditional distribution is given by:

$$p(\mathbf{y}^t|\mathcal{D}, \mathbf{x}^t, \mathcal{X}^\star, \mathcal{Y}^\star) = \int p(\mathbf{y}^t|\mathbf{b}^t, \mathbf{x}^t) p(\mathbf{b}^t, \mathbf{b}^T|\mathcal{D}, \mathcal{X}^\star, \mathcal{Y}^\star) d\mathbf{b}^t d\mathbf{b}^T \,, \tag{5}$$

where $\mathbf{b}^t$ are the noiseless black-box functions for fidelity t, including objectives and constraints, and $p(\mathbf{y}^t|\mathbf{b}^t, \mathbf{x}^t)$ are Gaussian likelihood factors to incorporate noise in the observations. The conditional distribution $p(\mathbf{b}^t, \mathbf{b}^T|\mathcal{D}, \mathcal{X}^\star, \mathcal{Y}^\star)$ is:

$$p(\mathbf{b}^t, \mathbf{b}^T|\mathcal{D}, \mathcal{X}^\star, \mathcal{Y}^\star) = Z^{-1} p(\mathbf{b}^t, \mathbf{b}^T|\mathcal{D}) p(\mathcal{X}^\star, \mathcal{Y}^\star|\mathbf{b}^T) \tag{6}$$

where Z is a normalization constant and we used that $(\mathcal{X}^\star, \mathcal{Y}^\star)$ only depends on fidelity T. $p(\mathbf{b}^t, \mathbf{b}^T|\mathcal{D})$ is the posterior for the black-boxes, and the term $p(\mathcal{X}^\star, \mathcal{Y}^\star|\mathbf{b}^T)$ is close to 0 for any $(\mathcal{X}^\star, \mathcal{Y}^\star)$ not fulfilling the constraints nor being Pareto-optimal, and larger than 0 and approximately constant otherwise:

$$p(\mathcal{X}^\star, \mathcal{Y}^\star|\mathbf{b}^T) \approx \prod_{\mathbf{x}^\star \in \mathcal{X}^\star} \left[\prod_{j=1}^{C} \tilde{\Theta}(c_j(\mathbf{x}^\star)) \right] \left[\prod_{\mathbf{x} \in \mathcal{X}} \tilde{\Omega}(\mathbf{x}, \mathbf{x}^\star) \right] \prod_{k=1}^{K} \mathcal{N}(y_k^\star(\mathbf{x}^\star)|f_k(\mathbf{x}^\star), \delta) \,, \tag{7}$$

where $\tilde{\Theta}(\cdot)$ is $1-\epsilon$ if $c_j(\mathbf{x}^\star) \geq 0$ and ϵ otherwise (*i.e.*, $\mathbf{x}^\star$ must be feasible); $\tilde{\Omega}(\cdot)$ is ϵ if $\mathbf{x}$ is feasible and dominates $\mathbf{x}^\star$ and $1-\epsilon$ otherwise (see [1]); $y_k^\star(\mathbf{x}^\star) \in \mathcal{Y}^\star$ is the Pareto front point for the k-th objective associated to $\mathbf{x}^\star$; finally δ is a small constant. In our implementation we set $\epsilon = 10^{-3}$. Moreover, the Gaussian factors in the r.h.s. of (7) guarantee that at $\mathcal{X}^\star$ we obtain the Pareto front $\mathcal{Y}^\star$.

Inspecting (6) we observe that it simply incorporates the extra factors in (7) into $p(\mathbf{y}^t|\mathcal{D}, \mathbf{x}^t)$. These factors can be easily included in the objective of each MF-DGP in (2), in the data-dependent term. The required expectations are tractable. Thus, $p(\mathbf{y}^t|\mathcal{D}, \mathbf{x}^t, \mathcal{X}^\star, \mathcal{Y}^\star)$ can be obtained by re-training the MF-DGP models simultaneously while incorporating the extra factors in (7). We approximate $\mathcal{X}$ in (7) by a random set of 100 instances generated uniformly when processing each mini-batch. After training, we also approximate the entropy of $p(\mathbf{y}^t|\mathcal{D}, \mathbf{x}^t, \mathcal{X}^\star, \mathcal{Y}^\star)$ by the entropy of a Gaussian with the same moments.

3 Experiments

We compare MF-JESMOC with other methods. Namely, MF-OSEMO [3], MOMF [4], MF-SEGO [5], MF-HVKG [6], and MF-CMOBO [7]. Only MF-CMOBO considers constraints (ignoring lower fidelities). We incorporate constraints in the other methods by multiplying the acquisition by the feasibility probability. MF-OSEMO and MF-SEGO use standard GPs to model each fidelity and assume linear dependencies. MOMF and MF-HVKG insert the fidelity level as an extra feature in a GP. This requires a lot of data to capture fidelity dependencies. MF-CMOBO models fidelities using a GP with prior mean given by the posterior

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

mean of the lower fidelity. Thus, lower-fidelity observations will not reduce the uncertainty of the higher fidelity. MF-JESMOC's acquisition is the only one that considers different fidelity levels for the constraints. Last, we also compare results with three variants of MF-JESMOC. Namely, RANDOM-HF, which evaluates the highest fidelity at random. MF-JESMOC-HF, which only evaluates the highest fidelity, and MF-JESMOC-LF which optimizes the lowest fidelity. The code for MF-JESMOC is found at https://github.com/fernandezdaniel/MOBOCMF.

We evaluate each method when finding an optimal ensemble of trees with minimum error on the German-Credit dataset and minimum size in terms of the number of nodes. We consider also, as a constraint, that the prediction cost is at most 7% of the original prediction cost when using a dynamic ensemble pruning technique. See [1] for further details. The parameters optimized are those described in [1], except for the ensemble size, which is used to obtain different fidelity levels. We consider two fidelities levels. The highest fidelity corresponds to an ensemble of 1000 trees and the lowest fidelity to an ensemble of 100 trees. Thus, $C_1 = 1$ and $C_2 = 10$, approximately. The evaluation budget is set equal to 50. We considered RBF kernels in all methods. When a method provides an infeasible point in the estimate of $\mathcal{X}^\star$, we ignore that point. The number of initial observations of the lower fidelity is 10 and 20 for the higher fidelity. We recommend by optimizing the high-fidelity posterior mean at each iteration, and guarantee that constraints are satisfied with 95% probability.

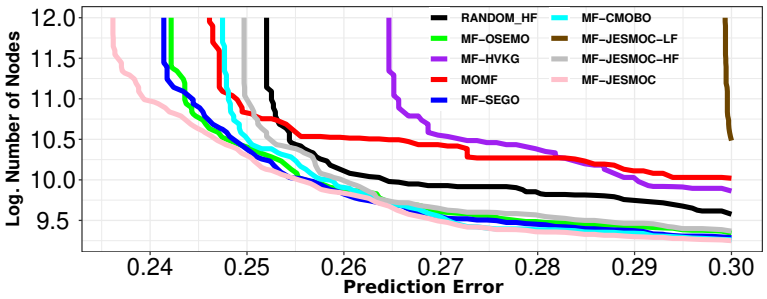

Fig. 2: Avg. Pareto front of each method when finding an optimal ensemble.

Fig. 2 shows the average Pareto-front of each method across 100 experiment repetitions. The best method is MF-JESMOC, which finds the most accurate ensembles with the smallest number of nodes. We observe that using low-fidelity data generally improves the results of only using high-fidelity data. MF-HVKG and MOMF perform bad because including the fidelity as a feature deteriorates the GP fit of the data. MF-CMOBO, MF-SEGO and MF-OSEMO improve over MF-JESMOC-HF, but do not perform as well as MF-JESMOC. Finally, MF-JESMOC-LF performs poorly since it solves a different optimization problem. Most low-fidelity ensembles are infeasible and do not satisfy the constraint.

Table 1 shows the median hyper-volume (higher better) of each method. We use the median instead of the average since it is more robust. Again MF-JESMOC gives the best results and the differences w.r.t. the other methods are statistically significant. Baseline results are not reported because lack of space.

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

MF-JESMOC recommends on average 15% of infeasible points. MF-OSEMO and MF-SEGO recommend on average 29% and 23% infeasible points, respectively.

Table 1: Median of the hyper-volumes returned by each method $\times 10$.

MF-JESMOC	MF-CMOBO	MF-OSEMO	MF-SEGO	MOMF	MF-HVKG
1.399±.005	1.223±.003	1.275±.002	1.321±.003	0.085±.004	0.073±.007

4 Conclusions

We have proposed MF-JESMOC, a strategy for Multi-objective BO with constraints and several fidelities. MF-JESMOC uses MF-DGPs to model individual black-boxes and the dependencies between different fidelities. By using the information provided by cheap low fidelities MF-JESMOC improves the optimization results w.r.t. optimizing directly the highest fidelity and w.r.t. other multi-fidelity methods from the literature. Specifically, it finds solutions with a higher hyper-volume than those of other methods and with a higher feasible probability.

References

[1] E. C. Garrido-Merchán and D. Hernández-Lobato. Predictive entropy search for multi-objective Bayesian optimization with constraints. *Neurocomputing*, 361:50–68, 2019.

[2] D. Fernández-Sánchez, E. C. Garrido-Merchán, and D. Hernández-Lobato. Improved max-value entropy search for multi-objective Bayesian optimization with constraints. *Neurocomputing*, 546:126290, 2023.

[3] S. Belakaria, A. Deshwal, and J. R. Doppa. Multi-fidelity multi-objective Bayesian optimization: An output space entropy search approach. In *Proceedings of the AAAI Conference on artificial intelligence*, pages 10035–10043, 2020.

[4] F. Irshad, S. Karsch, and A. Döpp. Expected hypervolume improvement for simultaneous multi-objective and multi-fidelity optimization. *arXiv preprint arXiv:2112.13901*, 2021.

[5] R. Charayron, T. Lefebvre, N. Bartoli, and J. Morlier. Towards a multi-fidelity & multi-objective Bayesian optimization efficient algorithm. *Aerospace Science and Technology*, 142:108673, 2023.

[6] S. Daulton, M. Balandat, and E. Bakshy. Hypervolume knowledge gradient: a lookahead approach for multi-objective Bayesian optimization with partial information. In *International Conference on Machine Learning*, 2023.

[7] Q. Lin, J. Hu, Q. Zhou, L. Shu, and A. Zhang. A multi-fidelity Bayesian optimization approach for constrained multi-objective optimization problems. *Journal of Mechanical Design*, 146:071702–1, 2024.

[8] K. Kandasamy, G. Dasarathy, J. Schneider, and B. Póczos. Multi-fidelity Bayesian optimisation with continuous approximations. In *International conference on machine learning*, pages 1799–1808. PMLR, 2017.

[9] H. Salimbeni and M. Deisenroth. Doubly stochastic variational inference for deep Gaussian processes. *Advances in neural information processing systems*, 2017.

[10] K. Cutajar, M. Pullin, A. Damianou, N. Lawrence, and J. González. Deep Gaussian processes for multi-fidelity modeling. In *NeurIPS Workshop on Bayesian Deep Learning*, 2018.

[11] C. Hvarfner, F. Hutter, and L. Nardi. Joint entropy search for maximally-informed Bayesian optimization. In *Advances in Neural Information Processing Systems*, 2022.

[12] J. M. Hernández-Lobato, M. W. Hoffman, and Z. Ghahramani. Predictive entropy search for efficient global optimization of black-box functions. *Advances in neural information processing systems*, pages 918–926, 2014.