ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

Transfer learning in Bayesian optimization for the calibration of a beam line in proton therapy

Valentin Hamaide* and François Glineur

UCLouvain - ICTEAM & CORE
Louvain-la-Neuve - Belgium

Abstract. Bayesian optimization (BO) is a type of black-box method used to optimize a costly objective function for which we have no access to derivatives. In practice, it is frequent that a series of similar problems has to be solved, with the problem data changing moderately between instances. We investigate a transfer learning approach based on BO that reuses information from a previous configuration in order to speed up subsequent optimizations. Our approach involves learning the noise variance to apply to the function values of the previous configuration and adapting the exploration-exploitation trade-off of the acquisition function from the previous configuration. We apply those ideas to the calibration of a beam line in proton therapy where the goal is to find magnet currents to obtain a desired shape for the beam of protons, and for which the calibration has to be repeated for several configurations. We show that reusing information from a previous configuration allows a reduction in the number of iterations by more than 80%, and that using BO is superior to the conventional Nelder-Mead algorithm for black box optimization and transfer learning.

1 Introduction

Solving a black box optimization problem without access to the gradient of the objective function can only be done in two ways. Either the gradient is estimated numerically and gradient-based algorithms can be used, or a derivative-free optimization routine is used. The former requires to compute a gradient at every step, which increases the number of function evaluations required. For costly evaluations of the objective function, this is a strong caveat. In this paper, we focus on Bayesian optimization, a derivative-free and global optimization routine [1] presented in section 2. For the sake of comparison with our approach, we also implement the Nelder-Mead algorithm, a conventional and simple derivative-free algorithm based on the concept of using the simplex as the iterate [2].

Transfer learning is usually described as transferring information from a related domain to improve the learner's capability, or because of insufficient data available for the actual learning task [3]. In this paper however, we consider that the transfer of information is from the same domain but from a different configuration. Transfer of information in the context of Bayesian optimization has been studied in the literature for tuning the hyperparameters of machine learning algorithms. In [4], the authors used a Gaussian process (GP) to learn a

*This work is supported by the Biowin-Bidmed project funded by the Walloon Region.

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

surrogate-based ranking function to transfer knowledge across tasks. In [5], the authors transfer knowledge from past experiments using deviations from the previous dataset mean via a common surrogate function. More recently, in [6], the authors introduced a noise variance to model the relatedness between datasets and estimate it via an inverse gamma distribution. In this paper, we propose a similar idea as in [6] but we estimate the noise by maximizing the log-marginal likelihood of the GP model. Moreover, we use mutual information [7] as the acquisition function of the BO and reuse the exploration-exploitation trade-off of the initial task to optimize the second, as explained in Section 3. In Section 4, we apply our method to the calibration of a beam line in proton therapy (PT).

2 Bayesian optimization

Bayesian optimization is a popular approach for solving black-box optimization problems. It is best-suited for problems of moderate dimension, and is able to handle noisy function evaluations [8]. The method works in two phases: first it builds a surrogate model of the objective function and quantifies the uncertainty in that surrogate via a Gaussian process regression. Then, it suggests the next point to evaluate by maximizing an acquisition function.

Gaussian process Gaussian processes (GP) are a form of non-parametric method that infer a distribution over functions by defining a prior [9]. After observing data points $D = (\mathbf{x}_i, y_i)_{i=1}^N$, it computes a posterior over functions. It relies on the property that a set of N points $\mathbf{x}_i$ induces a multivariate Gaussian distribution on $\mathbb{R}^N$. Formally, a Gaussian process is a random process where a point $\mathbf{x}_i$ is assigned a random variable $f(\mathbf{x}_i)$ and where the joint distribution of those variables $p(f(\mathbf{x}_1), ..., f(\mathbf{x}_N))$ is Gaussian, i.e. $p(\mathbf{f}|\mathbf{x}_1, ..., \mathbf{x}_N) = \mathcal{N}(\mathbf{0}, \mathbf{K})$ with $\mathbf{f} = [f(\mathbf{x}_1), ..., f(\mathbf{x}_N)]^T$ and $\mathbf{K}$ is a $N \times N$ kernel matrix with entries given by the kernel function (or covariance function) $K_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j)$, defining the shape of the function. Note that we consider the GP with a zero-mean function. This is common practice and not necessarily a limitation, since the mean of the posterior process is not confined to be zero [9]. To predict new values $\mathbf{f}_*$ at certain inputs $\mathbf{x}_*$, we compute a posterior distribution based on the partitioning

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{K} & \mathbf{K}_* \\ \mathbf{K}_*^T & \mathbf{K}_{**} \end{bmatrix}\right) \tag{1}$$

where $\mathbf{K}_* = \kappa(\mathbf{x}, \mathbf{x}_*)$ is a $N \times N_*$ matrix and $\mathbf{K}_{**} = \kappa(\mathbf{x}_*, \mathbf{x}_*)$ is a $N_* \times N_*$ matrix. Using rules for conditional probability on Gaussian distribution (see [9], section A.2 for details), we can compute the posterior distribution:

$$p(\mathbf{f}_*|\mathbf{X}_*, \mathbf{X}, \mathbf{f}) = \mathcal{N}(\mathbf{f}_*|\boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*) \text{ with } \boldsymbol{\mu_*} = \mathbf{K}_*^T\mathbf{K}^{-1}\mathbf{f} \text{ and } \boldsymbol{\Sigma_*} = \mathbf{K}_{**} - \mathbf{K}_*^T\mathbf{K}^{-1}\mathbf{K}_*$$

The Gaussian and the Matern kernel are commonly used, the latter being a generalization of the Gaussian kernel to relax its infinitely differentiable property [10]. We use a particular parametrization of the kernel that is twice differentiable

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \sigma_f^2 \left(1 + \frac{\sqrt{5}}{l}\|\mathbf{x}_i - \mathbf{x}_j\| + \frac{5}{3l}\|\mathbf{x}_i - \mathbf{x}_j\|^2\right) \exp\left(-\frac{\sqrt{5}}{l}\|\mathbf{x}_i - \mathbf{x}_j\|\right) \tag{2}$$

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

where σ_f^2 and l are parameters to tune. Those hyperparameters $\boldsymbol{\theta} = (\sigma_f^2, l)$ can be estimated optimally by maximizing the log marginal likelihood (LML) of the GP model [9], for which an analytical formulation exists:

$$\log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) = \log \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{K}_y, \boldsymbol{\theta}) = -\frac{1}{2}\mathbf{y}^T \mathbf{K}_n^{-1}(\boldsymbol{\theta})\mathbf{y} - \frac{1}{2}\log\left|\mathbf{K}_n(\boldsymbol{\theta})\right| - \frac{N}{2}\log(2\pi) \tag{3}$$

We can then compute the gradient of LML with respect to $\boldsymbol{\theta}$ and use an off-the-shelf optimization solver to obtain the optimal kernel parameters. In this paper, we use the L-BFGS-B algorithm [11] with several restarts due to the non-convexity of the problem, to perform all log-likelihood maximization. It is also possible to build a Gaussian process based on noisy evaluations ($\mathbf{y}|\mathbf{f} \sim \mathcal{N}(\mathbf{f}, \sigma_n^2\mathbf{I})$ where $\mathbf{y}$ are the observations) by defining the covariance matrix as $\mathbf{K}_n = \mathbf{K} + \sigma_n^2\mathbf{I}$ where σ_n^2 can be optimally retrieved by maximizing LML as well.

Bayesian optimization The next step after building a surrogate model of the objective function with GP is determining where to sample next by maximizing an acquisition function a. The acquisition function trades off between exploration of regions with high uncertainty and exploitation of regions with more knowledge. Formally, the next sample to evaluate $\mathbf{x}_{t+1}$ results from $\mathbf{x}_{t+1} = \arg\max_x a(\mathbf{x}|D_{1:t})$ with $D_{1:t}$, the dataset containing the t samples previously evaluated and used to build the surrogate model. The acquisition function we use in this paper is the mutual information, which was shown to surpass other common functions on several datasets [7]. It can be formulated as

$$a_{\mathrm{MI}}(\mathbf{x}) = \mu(\mathbf{x}) - \sqrt{\alpha}\left(\sqrt{\hat{\gamma}(t)} - \sqrt{\hat{\gamma}(t-1)}\right) \tag{4}$$

in the context of minimization, with $\mu(\mathbf{x})$ the mean at sample $\mathbf{x}$, representing the exploitation part, and the second term representing the exploration part at $\mathbf{x}$, where $\hat{\gamma}(t) = \sum_{i=1}^{t} \sigma_i^2(\mathbf{x})$ with $\sigma_i^2(\mathbf{x})$ the variance of the i^{th} GP at $\mathbf{x}$. This second term is empirically controlled by the amount of exploration that has been already done, that is, the more the algorithm has gathered information on f, the more it will focus on the optimum [7]. The parameter α controls the trade-off between precision and confidence.

3 Transfer learning approach

In this paper, we consider that an optimization problem is first solved in a source configuration with dataset D^S of size N_S. Then we wish to solve it again in a target configuration D^T of size N_T applying a transfer learning approach. The approach for transfer learning we propose in this paper is two-fold. First, we infer the noise variance of D^S with respect to D^T. We define $\sigma_{n_S}^2$ to be the noise variance of the source configuration with respect to the target configuration, and $\sigma_{n_T}^2$ to be the noise on the observations of the target configuration. We can take into account those noise variances directly in the GP by reformulating the

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

covariance matrix $\mathbf{K}$ as

$$\mathbf{K}_n = \mathbf{K} + \sigma_{n_S}^2 \left[\begin{array}{c|c} \mathbf{I}_{N_S} & \mathbf{0}_{N_S \times N_T} \\ \hline \mathbf{0}_{N_T \times N_S} & \mathbf{0}_{N_S \times N_T} \end{array}\right] + \sigma_{n_T}^2 \left[\begin{array}{c|c} \mathbf{0}_{N_S \times N_S} & \mathbf{0}_{N_S \times N_T} \\ \hline \mathbf{0}_{N_T \times N_S} & \mathbf{I}_{N_T} \end{array}\right] \quad (5)$$

The optimal noise variances can be estimated in the same way as the kernel hyperparameters, i.e. by maximizing the LML with respect to $\sigma_{n_S}^2$ and $\sigma_{n_T}^2$. Note that the approach can easily be extended to take into account more datasets by seeking different noise variances. However, the computation of the optimal parameters can quickly become time-consuming. A similar approach was undertaken in [6], in which the noise variance was estimated via an inverse gamma distribution instead of maximizing LML. We show in section 4, that the inverse gamma approach performs worse for our specific application. In this paper, we assume that the noise on the observations is fixed and very low. The purpose is thus to seek the optimal $\sigma_{n_S}^2$ for transfer learning, i.e. $\sigma_{n_S} = \arg\max_{\sigma_{n_S}} \log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})$.

The second ingredient in our approach is based on a modified quantity $\hat{\gamma}_T$ to empirically control the trade-off between exploration and exploitation of the source configuration. We define

$$\hat{\gamma}_T(t) = \sum_{j=1}^{t} \sigma_{T_j}^2(x_j) + C \sum_{i=1}^{N_S} \sigma_{S_i}^2(x_i) \quad (6)$$

where $\sigma_{T_j}^2$ and $\sigma_{S_i}^2$ are the variance of the j^{th} GP of the target and i^{th} GP of the source respectively, and $C \in [0, 1]$ to be a weight that we choose in this paper to be either $C = 1$ or $C = 0$ (although one could also adjust it w.r.t. $\sigma_{n_S}^2$).

4 Application to the calibration of a beam line in PT

We apply the approach developed in Section 3 to optimize magnet currents in a proton therapy beam line. Proton therapy (PT) is a type of radiotherapy that uses a beam of protons to irradiate cancerous tissues as opposed to a photon beam (X-rays) in conventional radiotherapy. A proton therapy system consists of a particle accelerator that accelerates protons, a beam line that transport the beam of protons to the treatment room and a delivery system responsible to irradiate a patient. Calibrating the magnet's currents in the beam line is necessary to fulfill several constraints on the characteristics of the beam of protons. The problem must be solved for several configurations, i.e. several beam ranges and gantry angles. We suppose the beam ranges to be independent problems, and we apply for each of them the transfer learning from one angle to another using the approach developed in section 3 to speed up this calibration process. The process is simulated via a fast tracking code for beam transport and simulation of beam-matter interactions in hadron therapy beamlines [12]. As input, we provide a list of magnet currents and receive a 2D dose distribution as output. An objective function translates this dose distribution into a scalar metric quantifying its correctness. The optimization stops when constraints on the beam are met, meaning that the correctness metric fell below some threshold, or when a maximum of 250 iterations has been reached.

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

Test with first angle First, we solve the optimization problem for the initial configuration, in our case a specific angle ($\theta_S = 0°$), for 10 beam ranges without transfer learning. We compare BO with the mutual information (BO-MI) acquisition function against the conventional lower confidence bound (BO-LCB) and the Nelder-Mead algorithm in the first part of Table 1. Experiments are restarted 20 times to account for the random nature of the BO. We observe that optimization of higher ranges is harder and the Nelder-Mead algorithm is not capable to find a solution satisfying the constraints in the maximal 250 iterations allowed. Between the two acquisition functions for the Bayesian optimization, MI performs better in almost all cases.

Test a second angle with transfer learning Next, we solve the calibration problem for another configuration ($\theta_T = 90°$) by reusing the currents-objective values pairs of the source configuration for each beam range. We also report a baseline without transfer learning for this new angle (second part of Table 1). We notice that the problem with this second angle seems easier for the Nelder-Mead algorithm, but is still failing for the first range (4.1) while BO-MI performs equally well on average and better than LCB. In the third part of the table, we compare the different approaches for transfer learning. We compare our proposed approach (σ_{MLE}, $C = 1$ in Table 1) with the approach from [6], i.e. by estimating the noise variance with an inverse gamma distribution, and with the Nelder-Mead algorithm where the initial simplex is built with the last values of the source configuration (the last $N + 1$ input-output pairs of the BO-MI for $\theta_S = 0$). From those results, we first observe that our approach of estimating $\sigma^2_{n_S} = \sigma^2_{MLE}$ by maximizing the LML of the GP and reusing the $\hat{\gamma}$ parameter of the MI acquisition function ($C = 1$) helps to reduce the number of iterations needed for the target configuration by more than 80% (ratio of the means). Second, it performs better than Nelder-Mead and the approach from [6]. Even if we modify the approach from [6] to select the same acquisition function as our approach (MI) and reuse the gamma parameter of the source configuration, we get worse results than σ_{MLE} (although better than the original method). This may be due to a poor prior candidate for the inverse gamma distribution. We also tested a few variations of our approach (σ_{MLE}). If we choose not to reuse the $\hat{\gamma}$ parameter of the MI acquisition function (i.e. $C = 0$), the number of iterations increases for all ranges. Moreover, if we arbitrarily fix the noise variance $\sigma = 0.1$ or $\sigma = 10^{-5}$ instead of estimating it, the number of iterations also increases.

5 Conclusion

We developed a transfer learning framework for Bayesian optimization based on the mutual information acquisition function that estimates the noise variance of a source task by maximizing the marginal likelihood of the Gaussian process and by reusing the exploration-exploitation trade-off of the source configuration. We showed that it drastically reduces the number of iterations needed to calibrate a proton therapy beam line and outperforms existing transfer learning approaches.

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

Method \ range	4.1	6	9	12	15	18	21	24	27	30	Mean
Source angle $\theta_S = 0°$											
BO-MI	85	47	32	18	32	47	58	69	114	108	61
BO-LCB	108	97	76	59	52	62	53	64	251	163	98.65
Nelder-Mead	33	7	12	42	251	251	251	251	251	251	160
Target angle $\theta_T = 90°$ without reusing information from θ_S											
BO-MI	102	64	22	16	28	60	40	65	97	96	59
BO-LCB	122	103	78	38	54	52	74	73	146	171	91.35
Nelder-Mead	251	45	16	11	46	57	52	59	56	56	64.9
Target angle $\theta_T = 90°$ reusing information from θ_S with BO-MI											
$\sigma_{MLE}, C = 1$	**21**	**12**	16	**5**	**6**	5	**4**	**3**	**5**	16	**9.3**
Nelder-Mead	24	28	**11**	20	9	10	10	8	6	**9**	13.5
Inv. gamma	194	122	74	63	77	97	68	143	194	121	115.3
Inv. gamma, MI	111	38	43	25	33	48	57	73	76	103	60.7
$\sigma_{MLE}, C = 0$	27	20	31	8	13	6	8	6	**5**	17	14.1
$\sigma = 10^{-5}, C = 1$	66	24	18	**5**	8	**4**	**4**	5	**5**	11	15
$\sigma = 0.1, C = 1$	79	53	32	27	39	30	52	64	52	50	47.8

Table 1: Median number of iterations to reach an acceptable solution for different transfer learning scenarios. Columns are different beam ranges (independent problems). The first part of the table refers to the source configuration ($\theta_S = 0°$), the second part refers to the target configuration ($\theta_T = 90°$) without transfer learning and the third part refers to the target configuration ($\theta_T = 90°$) and reuses the input-output-pairs computed by the best algorithm (BO-MI) of the first configuration ($\theta_S = 0°$).

References

[1] Jonas Mockus, Vytautas Tiesis, and Antanas Zilinskas. The application of bayesian methods for seeking the extremum. *Towards global optimization*, 2(117-129):2, 1978.

[2] John A Nelder and Roger Mead. A simplex method for function minimization. *The computer journal*, 7(4):308–313, 1965.

[3] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big data*, 3(1):1–40, 2016.

[4] Rémi Bardenet et al. Collaborative hyperparameter tuning. In *International conference on machine learning*, pages 199–207, 2013.

[5] Dani Yogatama and Gideon Mann. Efficient transfer learning method for automatic hyperparameter tuning. In *Artificial intelligence and statistics*, pages 1077–1085, 2014.

[6] Tinu Theckel Joy et al. A flexible transfer learning framework for bayesian optimization with convergence guarantee. *Expert Systems with Applications*, 115:656–672, 2019.

[7] Emile Contal et al. Gaussian process optimization with mutual information. In *International Conference on Machine Learning*, pages 253–261, 2014.

[8] 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, 2015.

[9] Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

[10] Michael L Stein. *Interpolation of spatial data: some theory for kriging*. Springer, 1999.

[11] Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1-3):503–528, 1989.

[12] Cedric Hernalsteens Robin Tesse, Kevin Andre. Optimization of hadron therapy beamlines using a novel fast tracking code for beam transport and beam-matter interactions. *International Conference on Atomic Physics*, 2018.