Hyperparameter Learning via Distributional Transfer

Ho Chung Leon Law
University of Oxford
ho.law@stats.ox.ac.uk

Peilin Zhao
Tencent AI Labs
masonzhao@tencent.com

Junzhou Huang
Tencent AI Labs
joehhuang@tencent.com

Dino Sejdinovic
University of Oxford
dino.sejdinovic@stats.ox.ac.uk

Abstract

Bayesian optimisation is a popular technique for hyperparameter learning but typically requires initial 'exploration' even in cases where potentially similar prior tasks have been solved. We propose to transfer information across tasks using kernel embeddings of distributions of training datasets used in those tasks. The resulting method has a faster convergence compared to existing baselines, in some cases requiring only a few evaluations of the target objective.

1 Introduction

Hyperparameter selection is an essential part of training a machine learning model and a judicious choice of values of hyperparameters such as learning rate, regularisation, or kernel parameters is what often makes a difference between an effective and a useless model. To tackle the challenge in a more principled way, the machine learning community has been increasingly focusing on Bayesian optimisation (BO) [12], a sequential strategy to select hyperparameters θ based on previous past evaluations of model performance. Using a Gaussian process (GP) [11] prior to build a representation of the underlying accuracy f as a function of the hyperparameters, and using an acquisition function $\alpha(\theta; f)$ to trade off exploration and exploitation, given the posterior of f, has been shown to give superior performance compared to traditional methods [12] such as grid search or random search. However, even in this setup, BO still suffers from the so called 'cold start' problem [10, 14]. Namely, in order to begin fitting a GP model, one needs to have initial evaluations of f at different hyperparameters. Hence, prior research considered transferring knowledge from previously 'solved' tasks, as in [14, 3, 13, 10]. However, to consider the similarity across tasks, initial random evaluations of the model at hand are often required. This might be prohibitive: evaluations of f can be computationally costly and our goal may be to select hyperparameters and deploy our model as soon as possible. We note that treating f as a black-box function, as is often the case in BO, is ignoring the highly structured nature of hyperparameter learning – it corresponds to training specific models on specific datasets. We make steps towards utilizing such structure in order to borrow strength across different tasks and datasets.

Contribution We assume a scenario where a number of tasks have already been 'solved' and propose a new BO algorithm, making use of the mean embeddings of the joint distribution of the training data [9, 2]. In particular, we propose a GP model that can jointly model all tasks at once, by considering extended domain of inputs to model accuracy f: joint distribution of the training data $\mathcal{P}_{XY}$, sample size of the training data s and hyperparameters θ. Through utilising all seen evaluations from all tasks and meta-information, we can optimise the marginal likelihood to learn a meaningful similarity between tasks. Experimentally, our methodology performs favourably already

Preprint. Work in progress.

at initialisation and has a faster convergence compared to existing baselines – in some cases, the optimal accuracy is achieved in just a few evaluations.

2 Related work

The idea of transferring information from different tasks in the context of BO is not new, and it has mainly been studied in the setting of multi-task BO [14, 3, 13, 10]. In multi-task BO, there is a set of tasks that we wish to solve jointly or we wish to solve a target task given some 'solved' source task. We follow the same setup, but take a different approach. Currently, correlation across tasks is only captured through evaluations of f. However, in terms of hyperparameter search for a machine learning model, this ignores additional information available: datasets used in training. Further, since task similarity is captured through evaluations of f only, this implies that we need to observe sufficient evaluations from the target task first in order to learn these task correlations. However, this is unnecessary in our proposed methodology, which can yield good initial hyperparameter candidates without having seen any evaluations from our target task, since we draw information from the meta-features corresponding to the training data distribution.

The use of such meta-information has in fact been explored before, but the current literature either uses hand-crafted manual features [4] or define them in an unsupervised way [6]. These strategies are not optimal, as while different tasks can have very different meta-information, their fs can in fact be highly correlated. In this case, using these meta-information (to define similarity) can have an adverse effect on exploration – this highlights the importance of using evaluations of f to learn the similarity between two tasks, as in multi-task BO. Furthermore, having obtained a similarity across tasks, current literature suggests to initialise with the best θ from the solved tasks, but again this is not optimal as we are neglecting non-optimal θ that can provide information for the target task. Our methodology can be seen as a combination of these two frameworks, as we use *learnt* embeddings of the joint distribution of the training data, while implicitly capturing correlation across tasks. The most similar in spirit to ours is the work of [7], who consider an additional input to be the sample size s, but do not consider different tasks corresponding to different training data distributions.

3 Background

Let f^{target} be the target task objective we would like to optimise, i.e. we want to find $\theta^*_{target} = \text{argmax}_{\theta\in\Theta} f^{target}(\theta)$. Assume that there are n (potentially) related source tasks f^i $i = 1, \dots n$. For each source task, we assume that we have $\{\theta^i_k, t^i_k\}_{k=1}^{N_i}$ from past runs, where t^i_k denotes a noisy evaluation of $f^i(\theta^i_k)$ and N_i denotes the number of evaluations of f^i from task i. Here, we assume that $f^i \; \forall \; i$ is the (underlying) accuracy of a trained machine learning model with training data $D_i = \{\mathbf{x}^i_l, y^i_l\}_{l=1}^{s_i}$, where $\mathbf{x}^i_l \in \mathbb{R}^p$ are the covariates, y^i_l are the labels and s_i is the size of the data. For a general framework, D is any input to f apart from θ – but following a typical supervised learning treatment, we assume it to be an iid sample from the joint distribution $\mathcal{P}_{XY}$. The method could be unsupervised though, as long as f is an appropriate form of a measure of performance. Under this setting, we have $(f^i, D_i = \{\mathbf{x}^i_l, y^i_l\}_{l=1}^{s_i}, \{\theta^i_k, t^i_k\}_{k=1}^{N_i})$ for $i = 1, \dots n$, where D_i denotes a dataset of source task i. Our strategy now is to measure the similarity between datasets (as a representation of the task itself), in order to find θ^*_{target}.

Assumptions To compare between different datasets, we make the assumption that $\mathbf{x}^i_l \in \mathcal{X}$ and $y^i_l \in \mathcal{Y}$ for all i, l, and that throughout the supervised learning model class M is the same. For example, one could imagine a data stream setting, where models have to be constantly retrained. This assumption implies that the source of differences of f^i across i and f^{target} is in the data D_i and D_{target} *only*, i.e. it is the dataset that affects the location of θ^*_i. In particular, we assume $D_i = \{\mathbf{x}^i_l, y^i_l\}_{l=1}^{s_i} \sim \mathcal{P}^i_{XY}$, where $\mathcal{P}^i_{XY}$ is the underlying joint distribution of the data for source task i. Further, as sample size is closely related to model complexity choice which is in turn closely related to hyperparameter choice [7], we will also encode this information.

With these assumptions, we consider $f(\theta, \mathcal{P}_{XY}, s)$, where f is a function on hyperparameters θ, joint distribution of the underlying data $\mathcal{P}_{XY}$ and sample size s. Here, f could be the negative of the empirical risk, i.e. $f(\theta, \mathcal{P}_{XY}, s) = -\frac{1}{s}\sum_{l=1}^{s} L(h_\theta(\mathbf{x}_l), y_l))$, where L is the loss function and h_θ is the model. In this form, we can also recover f^i and f^{target} from $f^i(\theta) = f(\theta, \mathcal{P}^i_{XY}, s_i)$ and

$f^{target}(\theta) = f(\theta, \mathcal{P}_{XY}^{target}, s_{target})$. Now by constructing an appropriate covariance function for these inputs, we can use a GP to model f. Intuitively, similarly to assuming that f varies smoothly as a function of θ in standard BO, this model also assumes smoothness of f across $\mathcal{P}_{XY}$[1] as well as across s following [7]. Intuitively, if two distributions and sample sizes are similar (in some distance that we will learn), their corresponding f will also be similar. In this source and target task setup, this would mean we can utilise information from *all* previous source datasets evaluations $\{\theta_k^i, t_k^i\}_{k=1}^{N_i}$.

4 Methodology

In order to model the quantities $\theta, \mathcal{P}_{XY}$ and s in a GP, we need a corresponding covariance function C on these quantities. Assuming a separable kernel for the inputs, we have:

$$C(\{\theta_1, \mathcal{P}_{XY}^1, s_1\}, \{\theta_2, \mathcal{P}_{XY}^2, s_2\}) = \nu k_\theta(\theta_1, \theta_2) k_p(\psi(D_1), \psi(D_2)) k_s(s_1, s_2) \tag{1}$$

where ν is a constant, k_θ and k_p is the standard Matérn-$3/2$ kernel in BO and k_s is the sample size kernel found in [7]. These are standard choices in the literature to facilitate fair comparisons, hence we do not investigate them further and instead we focus on modelling the dataset representation $\psi(D)$.

To specify $\psi(D)$, a feature map on joint distributions, estimated through samples D, we will follow an approach similar to [2] who consider transfer learning, and make use of kernel mean embeddings in order to compute feature maps of distributions (cf. [9] for an overview). In particular, we will consider feature maps of covariates and labels seperately, denoting them by $\phi_1(\mathbf{x}) \in \mathbb{R}^p$ and $\phi_2(y) \in \mathbb{R}^q$ [2]. Given these two feature maps, to meaningfully embed a joint distribution $\mathcal{P}_{XY}^i$, we use the cross covariance operator $\mathcal{C}_{XY}^i$ (see [5] for review), estimated by D_i with:

$$\hat{\mathcal{C}}_{XY}^i = \frac{1}{s_i} \sum_{\ell=1}^{s_i} \phi_1(\mathbf{x}_\ell^i) \otimes \phi_2(y_\ell^i), \tag{2}$$

i.e. in the case of finite-dimensional features, the cross covariance operator is just a tensor product of the feature maps $\Phi_1^i(\mathbf{x}) = [\phi_1(\mathbf{x}_1^i), \ldots, \phi_1(\mathbf{x}_{s_i}^i)] \in \mathbb{R}^{p \times s_i}$, $\Phi_2^i(y) = [\phi_2(y_1^i), \ldots, \phi_2(y_{s_i}^i)] \in \mathbb{R}^{q \times s_i}$,

$$\hat{\mathcal{C}}_{XY}^i = \frac{1}{s_i} \Phi_1^i(\mathbf{x}) \Phi_2^i(y)^\top \in \mathbb{R}^{p \times q}. \tag{3}$$

Given $\hat{\mathcal{C}}_{XY}^i$, we can flatten it to obtain $\psi(D_i) \in \mathbb{R}^{pq}$, an estimator of the representation of the joint distribution $\mathcal{P}_{XY}^i$. As $\psi(D_i) \in \mathbb{R}^{pq}$, we can use a kernel $k_p(s,t) : \mathbb{R}^{pq} \times \mathbb{R}^{pq} \to \mathbb{R}$ to measure similarity between $\mathcal{P}_{XY}^i$ and $\mathcal{P}_{XY}^j$ say. Note that while we have discussed the embedding of the joint distribution $\mathcal{P}_{XY}$, it is straightforward to also embed the marginal or conditional distributions.

An important choice is the form of $\phi_1(\mathbf{x})$ and $\phi_2(y)$, as these define the features of the distribution $\mathcal{P}_{XY}$ we would like to capture. For example $\phi_1(\mathbf{x}) = \mathbf{x}$ would be capturing the mean of the marginal distribution $\mathcal{P}_X$. In practice (keeping sample size s the same), if we know that $\mathcal{P}_{XY}^i \approx \mathcal{P}_{XY}^j$, we would expect that $\theta_i^* \approx \theta_j^*$, however generally the converse does not hold and the exact relationship is unknown. Hence, we opt for a flexible representation for $\phi_1(\mathbf{x})$ and $\phi_2(y)$ using small neural networks, which we optimise as part of the marginal likelihood maximisation. This can be thought of as a learning to learn setup [1], where a smaller robust model is used to optimise over bigger models. As we now have f on inputs $(\theta, \mathcal{P}_{XY}, s)$, we can fit a GP (with the standard normal noise model) on all observations $\{\{(\theta_k^i, \mathcal{P}_{XY}^i, s_i), t_k^i\}_{k=1}^{N_i}\}_{i=1}^n$ (along with any observations on the target), optimising any unknown parameters through the marginal likelihood. In order to propose the next θ^{target} to evaluate, we let $f^{target}(\theta) = f(\theta, \mathcal{P}_{XY}^{target}, s_{target})$ and maximise the acquisition function $\alpha(\theta; f^{target})$. Here, we use the expected improvement [8], however other options are readily applicable.

5 Experiments

We term the proposed method distBO and use a two layer neural network (of size 20 hidden units and 10 output units with RELU activation) for $\phi_1(\mathbf{x})$, while using an RBF network with 4 landmark

[1] We empirically verify validity of this assumption on simulated datasets in Appendix B.

[2] p and q can be potentially infinite, but we consider finite explicit feature maps for simplicity.

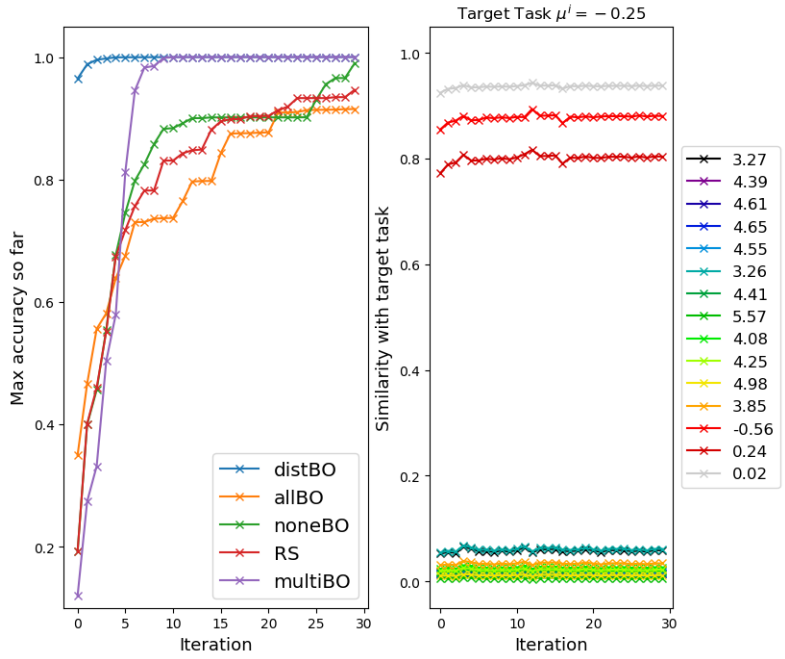

Figure 1: Target toy task with 15 source datasets **Left**: Max accuracy f seen so far **Right**: Similarity with the target task D_{target} for a particular run, measured using k_p. The legend represents the various sampled μ^i, here the target $\mu^i = -0.25$.

points for $\phi_2(y)$. In general, results are fairly robust to sensible settings of the neural network (that we *cannot* tune). For baselines, we will consider random search (RS), Bayesian optimisation (noneBO), multi-task Bayesian optimisation (multiBO) [14] (learns a similarity between tasks through evaluations of f *only*) and Bayesian optimisation with all source samples (allBO). The last baseline essentially ignores the distribution of the dataset, and assumes all seen samples come from the target task. For noneBO and multiBO, we will initialise with 5 iterations of random search.

We repeat each experiment 30 times and ignore k_s as samples sizes do not differ across task here. Throughout for both source and target tasks, we will run 30 iterations for each algorithm. To obtain $\{\theta_k^i, t_k^i\}_{k=1}^{30}$ for source task i, we use noneBO to simulate a realistic scenario.

Toy dataset To demonstrate our algorithm, we create 1-dimensional distribution of datasets from the following generative process:

$$\mu^i \sim \mathcal{N}(\gamma^i, 1) \quad X_l^i|\mu^i \sim \mathcal{N}(\mu^i, 1) \quad Y_l^i|X_l^i \sim \mathcal{N}(X_l^i, 1)$$

Given γ^i, we can simulate a μ^i as a characteristic varying across task, before sampling $D_i = \{x_l^i, y_l^i\}_{l=1}^{s_i}$ from $X_l^i, Y_l^i|\mu^i$. We consider a simple form of f given by

$$f(\theta; D_i) = \exp\left(-\frac{(\theta - \frac{1}{s}\sum_l x_l^i)^2}{2}\right),$$

and here $\theta \in [-8, 8]$ plays the role of a hyperparameter that we would like to learn and 'labels' y_l^i are just nuisance variables. The optimal hyperparameter is the sample mean of the $\{x_l^i\}$ and hence it is varying together with the underlying mean of the data μ^i. We now perform an experiment with $n = 15$, and $s_i = 400$, and generate 3 source task with $\gamma^i = 0$, and 12 source task with $\gamma^i = 4$. In addition, we generate an additional target dataset with $\gamma^i = 0$. The idea is that only 3 of the source tasks should be helpful to solve our target task, and distBO which can learn this should be able to reach the target optimum in only a few shots, as shown in Figure 1. Here, distBO is able to take advantage of the correct source tasks, allowing it to outperform other baselines. In particular this is evident on the right graph in Figure 1, which shows the similarity measure $k_p(\psi(D_i), \psi(D_{target})) \in [0, 1]$ for $i = 1, \ldots 15$. The feature representation has correctly learned to

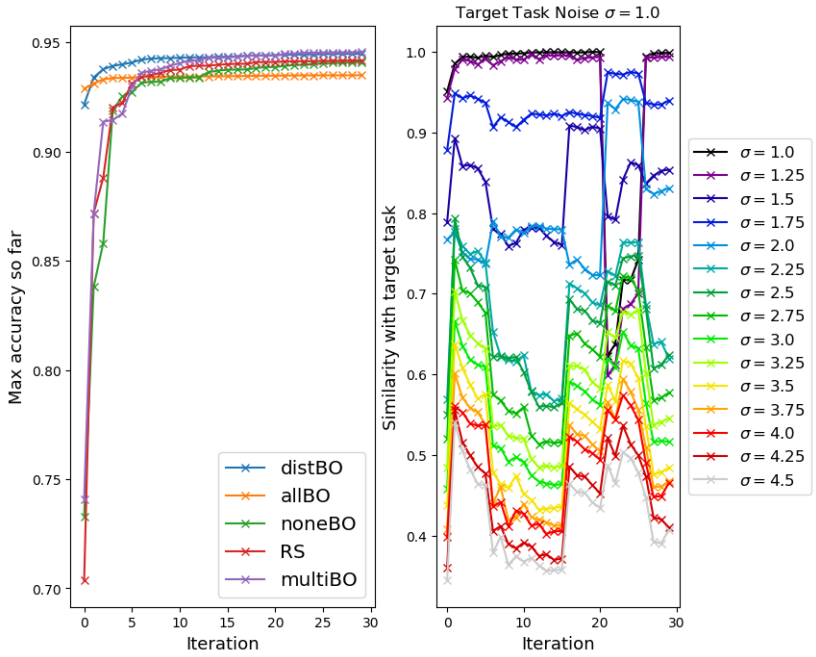

Figure 2: Target swiss rolls with $\sigma = 1.0$ with 15 source datasets **Left**: Max accuracy f seen so far on test set (Coefficient of Determinant, $R^2 \in [0, 1]$) **Right**: Similarity with the target task D_{target} for a particular run, measured using k_p. The legend represents source datasets, with various noise level σ. Here the target dataset has $\sigma = 1.0$.

put high similarity on the three source datasets from the same process and low similarity on source datasets that are not from the same process. We also observe that while the multi-task BO also achieve the target optimum quickly, it is unable to few-shot the optimum, as it does not make use of meta-information, hence needing intialisations from the target task to even begin learning the similarity across tasks. We also see that allBO which puts similarity as 1 for all tasks performs poorly.

Swiss Roll dataset We now consider the swiss roll dataset[3], as shown in Figure 4 in Appendix A, with covariates being the spatial location of each point, and labels given by the colour of the point. In order to make this problem more challenging, we place this data manifold in $\mathbb{R}^{10}$ by concatenating 7 dimensions of 0 onto the covariates, before applying a randomly selected, then fixed rotation matrix. We now fit a RBF kernel ridge regression (with hyperparameters α and γ), and aim to find the hyperparameters that will give us the highest accuracy as measured by the coefficient of determination (R^2). Throughout, we will take $s_i = 400$ for training and $s_i = 400$ for testing. In the first experiment, we have $n = 15$ source datasets with increasing noise on the marginal distribution of $\mathbf{x}$, with the target dataset having a noise of $\sigma = 1.0$. In figure 2, we observe that distBO can achieve the optimum in just several evaluations, this is non-surprising as distBO has correctly learnt the correct similarity ordering across tasks (with respect to the noise level).

Real dataset The Parkinson's disease telemonitoring dataset[4] consists of voice measurements using a telemonitoring device for 42 patients with Parkinson disease (200 recordings each). The label is the clinician's Parkinson disease symptom score for each recording. Following [2], we can treat each patient as a separate task. For the model, we employ RBF kernel ridge regression (with hyperparameters α, γ), with f as the coefficient of determination (R^2). While this problem does not gain from BO, as f is not expensive, this allows for a comprehensive benchmark comparisons. Here, we take a particular patient to be our target task, and randomly choose $n = 30$ other patients as source tasks (different across each repetition). We show the result in Figure 3 for three patients (as targets)

[3]The swiss roll manifold function (for sampling) can be found on the Python scikit-learn package, and there exists a variable σ for adding different noise levels.

[4]http://archive.ics.uci.edu/ml/datasets/Parkinsons+Telemonitoring

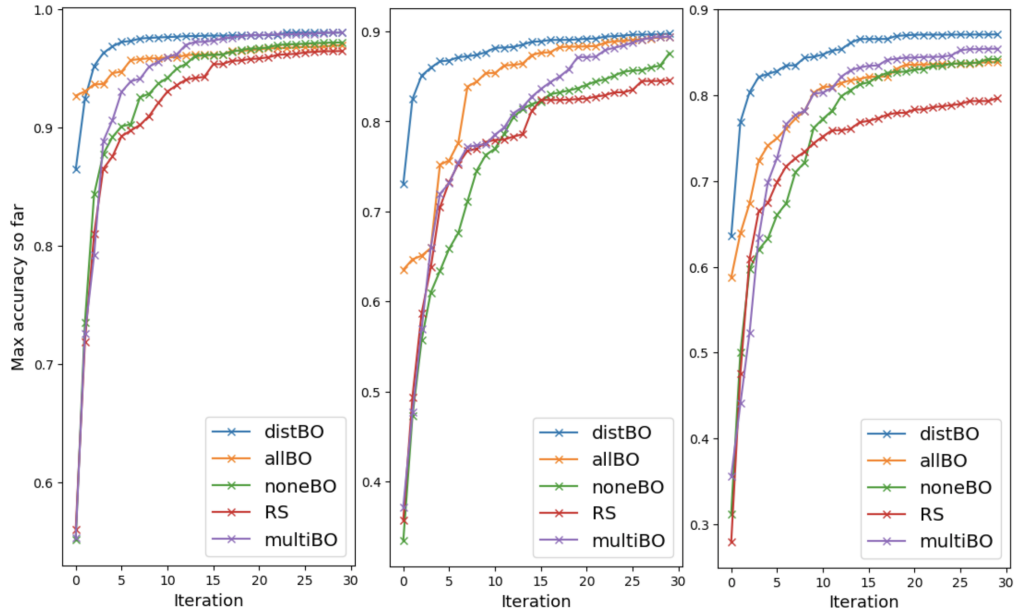

Figure 3: Max accuracy f seen so far on test set (Coefficient of Determinant, $R^2 \in [0, 1]$) on three different target patient task with 30 source datasets.

to highlight the behaviour. In fact across all patients being the target, distBO outperforms (or is equivalent to) all baselines in terms of faster convergence to the optimum. Here, it is clear that by encoding the distributional meta-information, we can learn useful similarities[5] between tasks.

6 Conclusion

We demonstrated that it is possible to borrow strength between multiple hyperparameter learning tasks by making use of the similarity between training datasets used in those tasks. This helped us to develop a method which finds a favourable setting of hyperparameters in only a few evaluations of the target objective. We argue that the model performance should not be treated as a black box function as it corresponds to specific known models and specific known datasets and that its careful consideration as a function of all its inputs, and not just of its hyperparameters, can lead to useful algorithms.

References

[1] Marcin Andrychowicz, Misha Denil, Sergio Gomez, Matthew W Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando De Freitas. Learning to learn by gradient descent by gradient descent. In *Advances in Neural Information Processing Systems*, pages 3981–3989, 2016.

[2] Gilles Blanchard, Aniket Anand Deshmukh, Urun Dogan, Gyemin Lee, and Clayton Scott. Domain generalization by marginal transfer learning. *arXiv preprint arXiv:1711.07910*, 2017.

[3] Matthias Feurer, Benjamin Letham, and Eytan Bakshy. Scalable meta-learning for bayesian optimization. *arXiv preprint arXiv:1802.02219*, 2018.

[4] Matthias Feurer, Jost Tobias Springenberg, and Frank Hutter. Initializing bayesian hyperparameter optimization via meta-learning. 2015.

[5] Arthur Gretton. Notes on mean embeddings and covariance operators. 2015.

[6] Jungtaek Kim, Saehoon Kim, and Seungjin Choi. Learning to transfer initializations for bayesian hyperparameter optimization. *arXiv preprint arXiv:1710.06219*, 2017.

[5] Shown in Figure 7 in Appendix C.

[7] Aaron Klein, Stefan Falkner, Simon Bartels, Philipp Hennig, and Frank Hutter. Fast bayesian optimization of machine learning hyperparameters on large datasets. *arXiv preprint arXiv:1605.07079*, 2016.

[8] J Močkus. On bayesian methods for seeking the extremum. In *Optimization Techniques IFIP Technical Conference*, pages 400–404. Springer, 1975.

[9] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, Bernhard Schölkopf, et al. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.

[10] Matthias Poloczek, Jialei Wang, and Peter I Frazier. Warm starting bayesian optimization. In *Proceedings of the 2016 Winter Simulation Conference*, pages 770–781. IEEE Press, 2016.

[11] Carl Edward Rasmussen. Gaussian processes in machine learning. In *Advanced lectures on machine learning*, pages 63–71. Springer, 2004.

[12] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959, 2012.

[13] Jost Tobias Springenberg, Aaron Klein, Stefan Falkner, and Frank Hutter. Bayesian optimization with robust bayesian neural networks. In *Advances in Neural Information Processing Systems*, pages 4134–4142, 2016.

[14] Kevin Swersky, Jasper Snoek, and Ryan P Adams. Multi-task bayesian optimization. In *Advances in neural information processing systems*, pages 2004–2012, 2013.

A Swiss Roll dataset

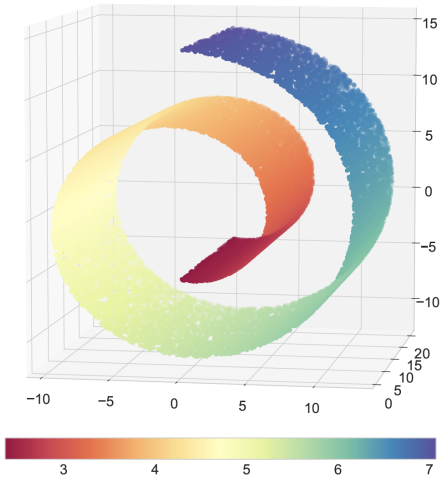

Figure 4: The swiss roll dataset, with no noise, i.e. $\sigma = 0$. The label of each point is given by its corresponding colour.

B Verifying Smoothness Assumption

We need to verify that as the joint distribution of the data changes, the change in the function f also changes smoothly. To provide some empirical results for this, we perform a grid search (using the same swiss roll setup in Section 5) to observe behaviour of f for increasing noise (changes in P_X) and increasing conditional distributional change (changes in $P_{Y|X}$ by taking a power P on label, before normalisation to [0,1] to obtain new label y_i^{new}), i.e.

$$y_i^{new} = \frac{(y_i)^{\mathrm{P}} - \min_i (y_i)^{\mathrm{P}}}{\max_i (y_i)^{\mathrm{P}} - \min_i (y_i)^{\mathrm{P}}} \tag{4}$$

The results are shown in Figure 5 and 6, and indeed we observe that f changes smoothly as the data distribution changes.

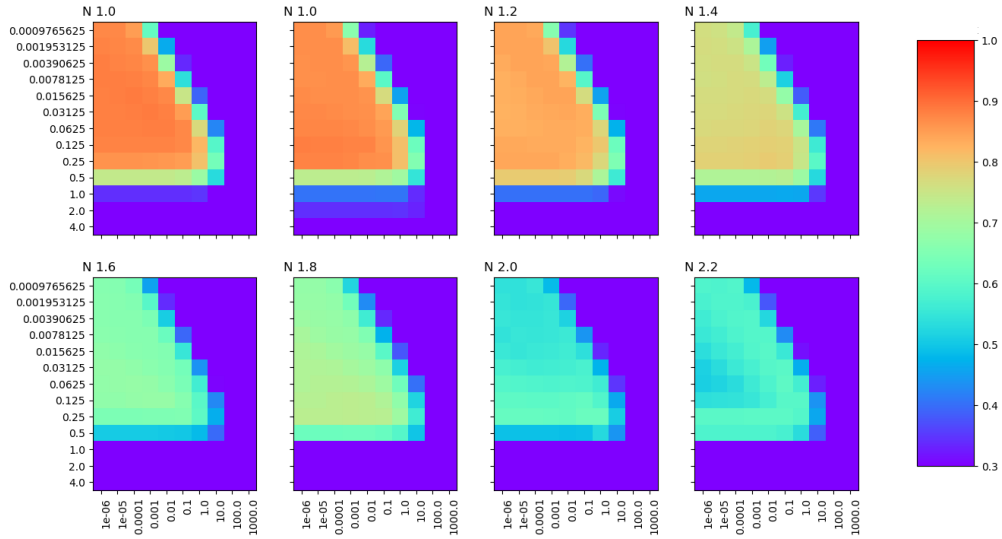

Figure 5: Swiss roll $R^2 \in [0, 1]$ on the test set for a grid of α, γ parameters for the kernel ridge regression. Here N denotes the additional noise σ on the covariates.

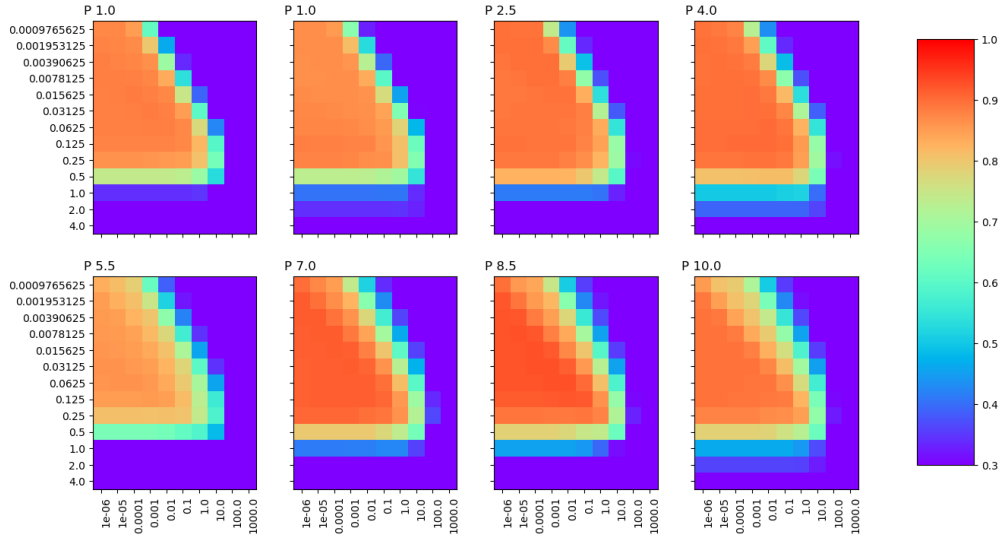

Figure 6: Swiss roll $R^2 \in [0, 1]$ on the test set for a grid of α, γ parameters for the kernel ridge regression. Here P denotes the additional power on the label (which we then normalise back to be in [0,1])

C Parkinson's disease dataset

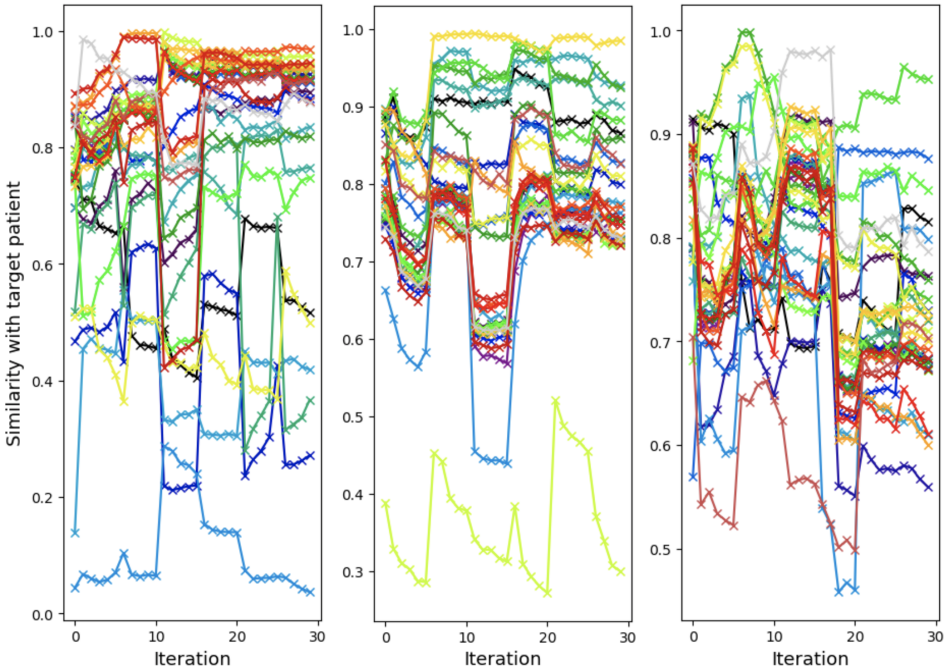

Figure 7: Similarities with the three different target patient for a particular repetition. Different colour represents a different patient source task.