Infinite-Task Learning with Vector-Valued RKHSs
[Talk submission]

Alex Lambert[1], Romain Brault[2], Zoltan Szabo[3], Maxime Sangnier[4], Florence d'Alché-Buc[1]

1: Télécom ParisTech - 2: Centrale-Supélec - 3: École Polytechnique - 4: Sorbonne Université

Abstract Machine learning has witnessed the success of solving tasks depending on a hyperparameter. While multi-task learning is celebrated for its capacity to solve jointly a finite number of tasks, learning a continuum of tasks for various loss functions is still a challenge. A promising approach, called *Parametric Task Learning*, has paved the way in the case of piecewise-linear loss functions. We propose a generic approach, called *Infinite Task Learning*, to solve jointly a continuum of tasks via Vector-Valued Reproducing Kernel Hilbert Spaces. We provide generalization guarantees to the suggested scheme and illustrate its efficiency in cost-sensitive classification, quantile regression and density level set estimation.

Keywords: operator-valued kernels, vv-rkhs, quantiles, functional learning

1 Motivation/Introduction

Several fundamental problems in machine learning and statistics can be phrased as the minimization of a loss function described by a hyperparameter. The hyperparameter might capture numerous aspects of the problem: (i) the tolerance w. r. t. outliers as the ϵ-insensitivity in SVR, (ii) importance of smoothness or sparsity such as the weight of the l_2-norm in Tikhonov regularization, l_1-norm in LASSO, (iii) Density Level-Set Estimation (DLSE), see for example one-class support vector machines One-Class Support Vector Machine (Schölkopf et al., 2000), (iv) confidence as examplified by Quantile Regression (QR, Koenker et al., 1978), or (v) importance of different decisions as implemented by Cost-Sensitive Classification (CSC, Zadrozny et al., 2001).

For some of these problems such as QR, CSC or DLSE, one is usually interested in solving the parametrized task for several hyperparameter values. When dealing with a finite number of those hyperparameters, multi-task learning (Evgeniou et al., 2004) is then a relevant setting, enabling to take benefit from the relationship between close parameterized tasks while keeping local properties of the algorithms: ν-property in DLSE (Glazer et al., 2013) or quantile property in QR (Takeuchi, Le, et al., 2006).

Eventually, it can be advantageous to allow the hyperparameter to change, possibly among infinitely many values in order to provide a prediction tool able to deal with any value of the hyperparameter. In their seminal work, (Takeuchi, Hongo, et al., 2013) extend the multi-task learning setting by considering an infinite number of parametrized tasks in a framework called Parametric Task Learning. They prove that under a piecewise-linearity assumption on the loss function, one recovers the task-wise solution for the whole spectrum of hyperparameters, at the cost of having a piecewise-linear model.

While being able to find the task-wise solution is a desired property, the strong assumption on the loss function and the restriction to a piecewise-linear model in the hyperparameter might be a hindrance. In this paper, we define a new family of tasks, called Infinite Task Learning, in which the piecewise linearity assumption on the loss is relaxed and whose goal is to learn a function with values in the space of continuous functions over the hyperparameter space. We propose to solve ITL in the context of vv-RKHS, shown to be adapted to multi-task learning (Micchelli et al., 2005). Due to space limitation, only the Quantile Regression problem is presented here.

2 The Infinite-Task learning framework

A *supervised parametrized task* is defined as follows. Let $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ be a random variable with joint distribution $\mathbf{P}_{X,Y}$; $\mathbf{P}_{X,Y}$ is assumed to be fixed but unknown. Instead we have access to n independent identically distributed observations called training samples: $\mathcal{S} := ((x_i, y_i))_{i=1}^{n} \sim \mathbf{P}_{X,Y}^{\otimes n}$. Let Θ be the domain of hyperparameters, and $\nu_\theta \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be a loss function associated to $\theta \in \Theta$. Let $\mathcal{H} \subset \mathcal{F}(\mathcal{X}; \mathcal{Y})$ denote our hypothesis class; the goal is to find a minimizer of the expected risk

$$R^{\theta}(h) := \mathbf{E}_{X,Y}[\nu_\theta(Y, h(X))], \tag{1}$$

QR*:* Assume $\mathcal{Y} \subseteq \mathbb{R}$ and $\theta \in [0, 1]$. For a given hyperparameter θ, Quantile Regression aims at predicting the θ-quantile of the real-valued output conditional distribution $\mathbf{P}_{Y|X}$. The task can be tackled (Koenker et al., 1978) using the pinball loss defined in Eq. (2).

$$\nu_\theta(y, h(x)) = |\theta - \mathbb{1}_{\mathbb{R}_-}(y - h(x))||y - h(x)| \tag{2}$$

The ITL framework aims at solving jointly a continuum of parametrized tasks. To that end, the following optimization problem is considered

$$\min_{h \in \mathcal{H}} R(h) := \mathbf{E}_{X,Y}\left[\int_\Theta \nu_\theta(Y, h(X)(\theta)) d\theta\right]. \tag{3}$$

Note that h is now a function-valued function, since at each point x we want a solution $h(x)$ to be able to predict a value at each hyperparameter. To modelize this, $\mathcal{H} \subset \mathcal{F}(\mathcal{X}; \mathcal{F}(\Theta; \mathcal{Y}))$ is chosen to be the vv-RKHS associated with the operator valued kernel $K(x, z) = k_{\mathcal{X}}(x, z) I_{\mathcal{H}_{k_\Theta}}$ where both $k_{\mathcal{X}}$ and k_Θ are classical scalar kernels, one on the input space, the other on the hyperparameter space. Since solving this problem without knowing $\mathbf{P}_{X,Y}$ is impossible, we rely on some empirical risk minimization strategy, along with a Quasi-Monte Carlo approximation with anchors $(\theta_j)_{j=1}^{m}$ to compute the integral. We also add a penalization based on the RKHS norm in $\mathcal{H}$, so that the problem to solve becomes

$$\arg\min_{h \in \mathcal{H}_K} \widetilde{R}_{\mathcal{S}}(h) + \frac{\lambda}{2}\|h\|_{\mathcal{H}_K}^2, \quad \lambda > 0. \tag{4}$$

where $\widetilde{R}_{\mathcal{S}}(h) := \frac{1}{nm} \sum_{i,j=1}^{n,m} \nu_{\theta_j}(y_i, h(x_i)(\theta_j))$.

3 Guarantees for the ITL scheme

Thanks to the choice of $\mathcal{H}$, Eq. (4) becomes amenable to optimization thanks to the following finite expansion.

Proposition 1 (Representer). *Assume that for* $\forall \theta \in \Theta, \nu_\theta$ *is a proper lower semicontinuous convex function with respect to its second argument. Then Eq. (4) has a unique solution* h^**, and* $\exists$ $(\alpha_{ij})_{i,j=1}^{n,m} \in \mathbb{R}^{n \times m}$ *such that* $\forall (x, \theta) \in \mathcal{X} \times \Theta$

$$h^*(x)(\theta) = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_{ij} k_{\mathcal{X}}(x, x_i) k_\Theta(\theta, \theta_j).$$

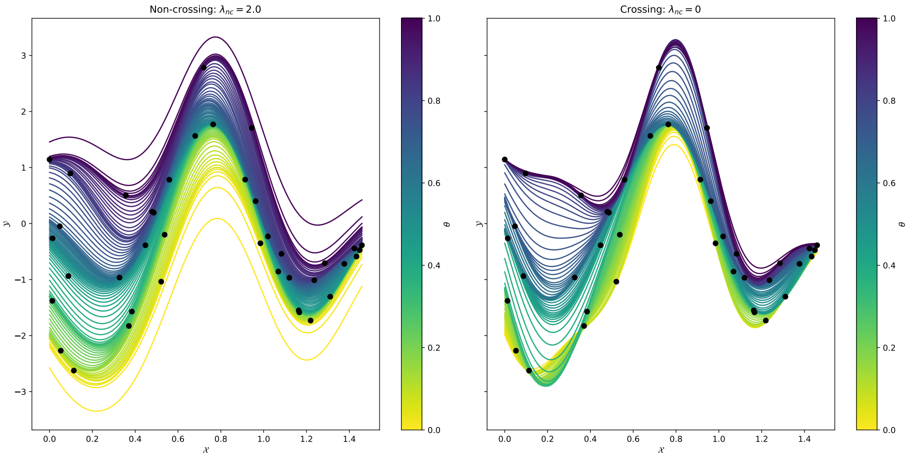

Figure 1: Impact of crossing penalty on toy data. Left plot: strong non-crossing penalty ($\lambda_{\text{nc}} = 2$). Right plot: no non-crossing penalty ($\lambda_{\text{nc}} = 0$). The plots show 100 quantiles of the continuum learned, linearly spaced between 0 (blue) and 1 (red).

In Prop. 2, we derive generalization error to the resulting estimate by stability argument (Bousquet et al., 2002), extending the work of Audiffren et al. (2013) to Infinite-Task Learning. We are especially interested in the effect of the two approximations, the one related to the size of the training sample and the other captured by m, the number of locations taken in the integral approximation. The key insight of Prop. 2 is that despite the two approximations (n, m), it is possible to get excess risk guarantees, highlighting the role of m and n.

Proposition 2 (Generalization). *Let $h^* \in \mathcal{H}$ be the solution of Eq. (4) for the* QR *or* CSC *problem with Quasi Monte Carlo approximation. Under mild conditions on the kernels $k_{\mathcal{X}}$, k_{Θ} and $\mathbf{P}_{X,Y}$, one has*

$$R(h^*) \leqslant \widetilde{R}_{\mathcal{S}}(h^*) + \mathcal{O}_{\mathbf{P}_{X,Y}}\left(\frac{1}{\sqrt{n}}\right) + \mathcal{O}\left(\frac{\log(m)}{m}\right).$$

Concerning the implementation of ITL, the optimization is performed on the $(\alpha_{ij})_{i,j=1}^{n,m}$ vector of size nm with L-BFGS on a smoothed version of the pinball loss. Moreover, having a continuous model in the hyperparameter allows us to design new penalty to enforce the non-crossing phenomenon between quantiles, namely

$$\widetilde{\Omega}_{\text{nc}}(h) = \frac{\lambda_{nc}}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} \left| -\frac{\partial h}{\partial \theta}(x_i)(\theta_j) \right|_{+} \tag{5}$$

The Fig. 1 illustrates the efficiency of this new constraint made possible by the continuum scheme.

4 Discussion/Conclusion

Infinite Task Learning with vv-RKHS is a novel nonparametric framework aiming at jointly solving parametrized tasks for a continuum of hyperparameters. This approach allows to recover several existing multi-task approaches and extends Parametric-Task Learning to nonparametric models and a larger class of loss functions.

References

Audiffren, Julien and Hachem Kadri (2013). “Stability of Multi-Task Kernel Regression Algorithms.” In: *Asian Conference on Machine Learning (ACML)*. Vol. 29. PMLR, pp. 1–16.

Bousquet, Olivier and André Elisseeff (2002). “Stability and generalization.” In: *Journal of Machine Learning Research* 2, pp. 499–526.

Evgeniou, Theodoros and Massimiliano Pontil (2004). “Regularized multi–task learning.” In: *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 109–117.

Glazer, Assaf, Michael Lindenbaum, and Shaul Markovitch (2013). “q-OCSVM: A q-quantile estimator for high-dimensional distributions.” In: *Advances in Neural Information Processing Systems (NIPS)*, pp. 503–511.

Koenker, Roger and Gilbert Bassett Jr (1978). “Regression quantiles.” In: *Econometrica: journal of the Econometric Society*, pp. 33–50.

Micchelli, Charles A. and Massimiliano Pontil (2005). “On Learning Vector-Valued Functions.” In: *Neural Computation* 17, pp. 177–204.

Schölkopf, Bernhard et al. (2000). “New support vector algorithms.” In: *Neural computation* 12.5, pp. 1207–1245.

Takeuchi, Ichiro, Tatsuya Hongo, et al. (2013). “Parametric task learning.” In: *Advances in Neural Information Processing Systems (NIPS)*, pp. 1358–1366.

Takeuchi, Ichiro, Quoc V Le, et al. (2006). “Nonparametric quantile estimation.” In: *Journal of Machine Learning Research* 7, pp. 1231–1264.

Zadrozny, Bianca and Charles Elkan (2001). “Learning and making decisions when costs and probabilities are both unknown.” In: *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 204–213.