Multiple-Kernel Regression with Sparsity Constraints

Shayan Aziznejad and Michael Unser
Biomedical Imaging Group, EPFL, Lausanne, Switzerland
Emails: shayan.aziznejad@epfl.ch, michael.unser@epfl.ch

***Abstract*—We consider the problem of learning a function from a sequence of its noisy samples in a continuous-domain hybrid search space. We adopt the generalized total-variation norm as a sparsity-promoting regularization term to make the problem well-posed. We prove that the solution of this problem admits a sparse kernel expansion with adaptive positions. We also show that the sparsity of the solution is upper-bounded by the number of data points. This allows for an enlargement of the search space and ensures the well-posedness of the problem.**

I. Introduction

The goal of supervised learning is to learn an unknown function $f:\mathbb{R}^d \to \mathbb{R}$ from a set of its noisy measurements $(\boldsymbol{x}_m, y_m)$, where $y_m \approx f(\boldsymbol{x}_m)$ for $m = 1, 2, \dots, M$. In a reproducing-kernel Hilbert space $\mathcal{H}(\mathbb{R}^d)$, this problem is commonly formulated through the minimization

$$\min_{f\in\mathcal{H}(\mathbb{R}^d)} \sum_{m=1}^{M} \left(f(\boldsymbol{x}_m) - y_m\right)^2 + \lambda \|f\|_{\mathcal{H}}^2. \tag{1}$$

It is known that the solution of (1) lies in the linear span of $\{\mathrm{k}(\cdot, \boldsymbol{x}_m)\}_{m=1}^M$, where $\mathrm{k}(\cdot,\cdot)$ is the unique reproducing kernel of $\mathcal{H}(\mathbb{R}^d)$ [1]. The form of the solution is then useful to reduce the continuous-domain minimization problem (1) to a discrete finite-dimensional problem that has a closed-form solution [2].

The multiple-kernel learning framework was proposed as a generalization of the classical method with the aim of increasing the model flexibility [3] [4]. In this approach, the kernel function itself is learned as a linear combination of some basis kernels.

II. Proposed Framework

The Schwartz space of smooth and rapidly decaying functions is denoted by $\mathcal{S}(\mathbb{R}^d)$. Its topological dual $\mathcal{S}'(\mathbb{R}^d)$ is the space of tempered distributions. An invertible linear shift-invariant operator L with the frequency response $\widehat{\mathrm{L}}(\omega)$ is called admissible if, for any $\varphi \in \mathcal{S}'(\mathbb{R}^d)$, $\mathrm{L}\{\varphi\} = \mathcal{F}^{-1}\{\widehat{\mathrm{L}}\widehat{\varphi}\}$ and $\mathrm{L}^{-1}\{\varphi\} = \mathcal{F}^{-1}\{\frac{\widehat{\varphi}}{\widehat{\mathrm{L}}}\}$ are both elements of $\mathcal{S}'(\mathbb{R}^d)$. The underlying kernel of L is then defined as $\mathrm{k} = \mathcal{F}^{-1}\{\frac{1}{\widehat{\mathrm{L}}}\} \in \mathcal{S}'(\mathbb{R}^d)$.

We follow the Banach-space framework of Unser *et al.* in [5] by imposing a sparsity-promoting regularization term called the generalized total variation (gTV). Given an admissible operator $\mathrm{L}: \mathcal{S}'(\mathbb{R}^d) \to \mathcal{S}'(\mathbb{R}^d)$, the gTV norm is defined as

$$\mathrm{gTV}(w) = \|\mathrm{L}\{w\}\|_{\mathcal{M}} \triangleq \sup_{\substack{\varphi\in\mathcal{S}(\mathbb{R}^d)\\ \|\varphi\|_\infty \le 1}} |\langle \mathrm{L}\{w\}, \varphi\rangle|. \tag{2}$$

The native space for the operator L is the Banach space of elements of $\mathcal{S}'(\mathbb{R}^d)$ with finite gTV norm, defined by

$$\mathcal{M}_{\mathrm{L}}(\mathbb{R}^d) = \{w \in \mathcal{S}'(\mathbb{R}^d) : \|\mathrm{L}\{w\}\|_{\mathcal{M}} < +\infty\}. \tag{3}$$

We propose a new multicomponent model for the target function f. We assume that $f = \sum_{n=1}^N f_n$, with $f_n \in \mathcal{M}_{\mathrm{L}_n}(\mathbb{R}^d)$, where each component f_n has a certain degree of smoothness in accordance with its corresponding regularization operator L_n. We impose our model priors through the minimization

$$\min_{\substack{f_n\in\mathcal{M}_{\mathrm{L}_n}(\mathbb{R}^d)\\ f=\sum_{n=1}^N f_n}} \sum_{m=1}^{M} \left(f(\boldsymbol{x}_m) - y_m\right)^2 + \lambda \sum_{n=1}^{N} \|\mathrm{L}_n\{f_n\}\|_{\mathcal{M}}. \tag{4}$$

Theorem 1 describes the solution form of (4).

Theorem 1. *There exists a solution* $(f_1, f_2, \dots, f_N)$ *of* (4) *such that the reconstructed function* $f = \sum_{n=1}^N f_n$ *takes the form*

$$f(\cdot) = \sum_{n=1}^{N}\sum_{j=1}^{M_n} a_{n,j}\mathrm{k}_n(\cdot - \boldsymbol{z}_{n,j}) \tag{5}$$

for some sparse coefficients $a_{n,j} \in \mathbb{R}$ *and adaptive positions* $\boldsymbol{z}_{n,j} \in \mathbb{R}^d$*. Moreover,* $\sum_{n=1}^N M_n \le M$ *and* $\sum_{n=1}^N \|\mathrm{L}_n\{f_n\}\|_{\mathcal{M}} = \sum_{n=1}^N \sum_{j=1}^{M_n} |a_{n,j}|$.

Theorem 1 has been proven in the extended version of this work [6]. It proposes an adaptive kernel expansion for the multiple-kernel regression model. The total number of active kernels (with nonzero coefficients) is upper-bounded by M and does not depend on the number of search spaces. This allows one to enlarge the search space while keeping the problem well-posed and nonreduntant. The gTV regularization also enforces an ℓ_1-penalty on the kernel coefficients, which results in a sparse kernel expansion.

III. Admissible Kernels

An important aspect of our theory is to identify the class of admissible kernels. We show that, for any function $\mathrm{k}: \mathbb{R}^d \to \mathbb{R}$, if $\widehat{\mathrm{k}}(\omega)$ and $\frac{1}{\widehat{\mathrm{k}}(\omega)}$ are smooth and slowly growing functions, then $\mathrm{k}(\cdot)$ is an admissible kernel. An example is the sub-Gaussian kernels defined as

$$\mathrm{k}_\alpha(\boldsymbol{x}, \boldsymbol{y}) = \exp(-\|\boldsymbol{x} - \boldsymbol{y}\|_\alpha^\alpha). \tag{6}$$

The tuning parameter $\alpha \in (0, 2)$ is related to the asymptotic decay of the kernel function in the Fourier domain. The case $\alpha = 2$ (Gaussian kernels) is excluded from our theory since the frequency response of the corresponding operator has exponential growth and, hence, is not in $\mathcal{S}'(\mathbb{R}^d)$. However, we can get arbitrarily close by letting $\alpha = (2 - \epsilon)$ for a small value of $\epsilon > 0$.

References

[1] G. Wahba, *Spline Models for Observational Data*. SIAM, 1990, vol. 59.
[2] B. Schölkopf and A. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2001.
[3] G. Lanckriet, N. Cristianini, P. Bartlett, L. Ghaoui, and M. Jordan, "Learning the kernel matrix with semidefinite programming," *Journal of Machine Learning Research*, vol. 5, pp. 27–72, Jan 2004.
[4] F. R. Bach, G. Lanckriet, and M. Jordan, "Multiple kernel learning, conic duality, and the SMO algorithm," in *Proceedings of the Twenty-First International Conference on Machine Learning*. ACM, 2004, p. 6.
[5] M. Unser, J. Fageot, and J. Ward, "Splines are universal solutions of linear inverse problems with generalized TV regularization," *SIAM Review*, vol. 59, no. 4, pp. 769–793, 2017.
[6] S. Aziznejad and M. Unser, "An L1 representer theorem for multiple-kernel regression," *arXiv preprint arXiv:1811.00836*, 2018.

This work was funded by the Swiss National Science Foundation under Grant 200020_184646 / 1.