Enhancing Kernel Flexibility via Learning Asymmetric Locally-Adaptive Kernels

Anonymous Authors[1]

Abstract

Insufficient flexibility poses a critical bottleneck for kernel-based learning that relies on manually designed and non-trainable kernels. This paper addresses this limitation by introducing an innovative solution: the asymmetric variant of Radial Basis Function (RBF) kernels and the corresponding kernel learning algorithm. These kernels incorporate Locally-Adaptive-Bandwidths (LAB) as trainable parameters, named as LAB RBF kernels, significantly enhancing kernel flexibility. To tackle challenges arising from asymmetry, we pioneer an asymmetric kernel ridge regression framework. We offer new theoretical insights into our algorithm's connection with an ℓ_0-regularized model in an integral space of Reproducible Kernel Hilbert Spaces. This analysis shows that by enhancing kernel flexibility, our approach widens the hypothesis space. This expansion helps us identify functions that fit the data better while ensuring strong generalization, as confirmed by numerical experiments on real datasets.

1. Introduction

Kernel methods play a foundational role within the machine learning community, offering a lot of classical non-linear algorithms, including Kernel Ridge Regression (KRR, (Vovk, 2013)), Support Vector Machines (SVM, (Cortes & Vapnik, 1995)), and a host of other innovative algorithms. Nowadays, kernel methods maintain their importance thanks to their interpretability, strong theoretical foundations, and versatility in handling diverse data types (Ghorbani et al., 2020; Bach, 2022; Jerbi et al., 2023). However, as newer techniques like deep learning gain prominence, kernel methods reveal a shortcoming: the learned function's flexibility often falls short of expectations.

A sufficiently flexible model, often characterized by over-parameterization (Allen-Zhu et al., 2019; Zhou & Huo, 2024), has attracted researchers' attention due to the phenomenon of benign overfitting: the ability to interpolate samples while maintaining good generalization (Ma et al., 2017; Montanari & Zhong, 2020). However, interpolation achieved using either single or multi-kernel methods typically relies on kernels close to the Dirac function and ridgeless models, resulting in large parameter norms and poor generalization. The imperfect interpolation observed in kernel-based learning can be attributed to its inherent lack of flexibility, characterized by a limited number of free parameters that fall short of the capabilities observed in over-parameterized deep models. Augmenting the trainable parameters of traditional kernel methods poses a primary challenge due to the reliance on manually designed, fixed kernels, which are inherently untrainable. Recognizing this problem, researchers have proposed kernel learning techniques such as multiple kernel learning (Liu et al., 2023b; Liu, 2023) and deep kernel learning (Wilson et al., 2016; Huang et al., 2023). The success of these approaches demonstrates the effectiveness of increasing kernel flexibility.

In this paper, we focus on kernel ridge regression problem and present a novel approach to enhance kernel flexibility through asymmetric kernel learning. We achieve this by introducing a variant of Radial Basis Function (RBF) kernels with trainable bandwidths. Given a dataset $\mathcal{X} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\} \subset \mathbb{R}^M$ and let $\odot$ denote the Hadamard product, then the kernel defined on $\mathcal{X}$ is outlined below: for $\forall \boldsymbol{x}_i \in \mathcal{X}, \forall \boldsymbol{t} \in \mathbb{R}^M$,

$$\mathcal{K}_{\boldsymbol{\Theta}}(\boldsymbol{t}, \boldsymbol{x}_i) = \exp\left\{-\|\boldsymbol{\theta}_i \odot (\boldsymbol{t} - \boldsymbol{x}_i)\|_2^2\right\}, \tag{1}$$

where $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_N\} \subset \mathbb{R}_+^M$ denotes a bandwidth set. We name (1) as Local-Adaptive-Bandwidth RBF (LAB RBF) kernels. The key difference between LAB RBF kernels and conventional RBF kernels lies in assigning distinct bandwidths $\boldsymbol{\theta}_i$ to each sample $\boldsymbol{x}_i$ rather than using a uniform bandwidth across all data points[1], and we propose to estimate these bandwidths $\boldsymbol{\theta}_i$ from training data.

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

[1] Strictly speaking, the bandwidth in LAB RBF kernels should be a vector function defined on $\mathbb{R}_+^M$. But for better clarity in the subsequent learning algorithm, we discretely define the bandwidth for each support vector data point in a point by point way.

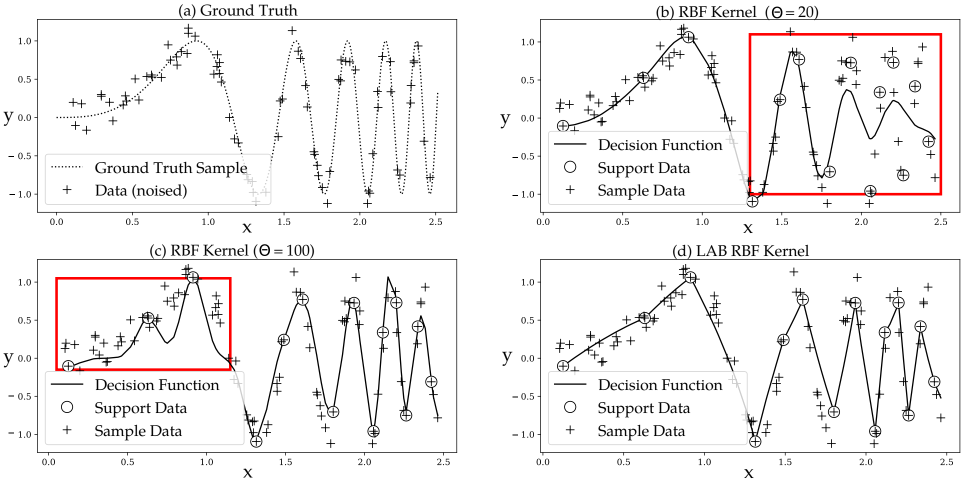

Figure 1. A toy example demonstrating the regression of a 1-dimensional signal $y = \sin(2x^3)$. (a) Ground truth. Functions obtained with a universal kernel bandwidth exhibit limitations: (b) lacks accuracy in high-frequency regions, while (c) results in unnecessary sharp changes in low-frequency areas. (d) By introducing locally adaptive bandwidths, our method achieves larger bandwidths on the left and smaller bandwidths on the right.

Obviously, the introduction of such data-dependent bandwidths can effectively increase the flexibility of kernel-based learning, making bandwidths adaptive to data. Figure 1 illustrates the advantages of incorporating locally-adaptive bandwidths. Traditional RBF kernel functions utilize a uniform bandwidth, which can present challenges in fitting signals with varying frequencies, such as $y = \sin(2x^3)$ in this toy example. When using a small bandwidth, the resulting function appears smooth but fails to capture the high-frequency portions, as highlighted in the red box in Fig.1(b). Conversely, employing a large bandwidth leads to an overly sharp function that inadequately represents the smooth portions, as highlighted in the red box in Fig.1(c). By allowing for different bandwidths in local regions, the proposed LAB RBF kernel emerges as an optimal solution, as depicted in Fig. 1(d).

Notably, LAB RBF kernels exhibit inherent asymmetry since $\boldsymbol{\theta}_i$ is not required to be equal to $\boldsymbol{\theta}_j$, allowing for cases where $\mathcal{K}(\boldsymbol{x}_i, \boldsymbol{x}_j) \neq \mathcal{K}(\boldsymbol{x}_j, \boldsymbol{x}_i)$. This asymmetry, while providing added flexibility to LAB RBF kernels compared to traditional symmetric kernels, presents both algorithmic and theoretical challenges: *how to incorporate asymmetric LAB RBF kernels into existing symmetric-kernel-based regression models? how to determine the optimal bandwidths?* and *what are the theoretical benefits in using the asymmetric LAB RBF kernels?* This paper addresses these challenges and makes the following contributions:

Asymmetric kernel ridge regression framework. This paper for the first time introduces an asymmetric KRR framework for the utilization of asymmetric LAB RBF kernels, presented from both primal and dual perspectives. An analytical expression for the stationary points is derived, which remains consistent with that of classical symmetric KRR model, despite the asymmetric nature of the kernel matrix.

Robust kernel learning algorithm. A novel kernel learning algorithm tailored for LAB RBF kernels is introduced, enabling the determination of local bandwidths. Empowered by the flexibility of LAB RBF kernels, this algorithm enables the regression function to effectively accommodate a diverse range of data patterns while maintaining a highly compact form.

Theoretical interpretation of LAB RBF kernels. This paper is the first to uncover the hypothesis space and inherent sparsity of LAB RBF kernels. Subsequently, we establish a connection between our algorithm and an ℓ_0-related model within an integral space of Reproducible Kernel Hilbert Spaces (RKHSs). Theoretical results demonstrate that employing asymmetric kernel learning to enhance kernel flexibility essentially enlarges the hypothesis space. Consequently, we are able to find a satisfactory function that is sparse while still achieving a good fit to the data.

The experimental results underscore the advanced performance of our algorithm: it achieves state-of-the-art regression accuracy while substantially reducing the required amount of support data (model complexity) compared to advanced kernel-based methods.

2. Related Works

Kernel ridge regression. The focus of this paper is on Kernel Ridge Regression (Vovk, 2013), a powerful regression technique that combines ridge regression with the kernel trick to model complex relationships in data. Due to its solid theoretical foundation and straightforward algorithm, KRR and its variants have continued to be widely studied in the machine learning community. Advanced studies have explored various applications, such as distributed datasets (Lin et al., 2020), graph data (Liu et al., 2023a), neural tangent kernels (Jacot-Guillarmod, 2022), and so on. However, the lack of a theoretical explanation hampers the use of asymmetric kernels in KRR. In this paper, we address this limitation by introducing an innovative asymmetric KRR framework from the primal-dual viewpoint.

RBF kernels with diverse bandwidths. Inspired by real-world instances of asymmetric similarity, the concept of asymmetric kernels, including LAB RBF kernels, has been in existence for a long time. In statistics, particularly in kernel regression and kernel density estimation, such as (Abramson, 1982; Brockmann et al., 1993; Mackenzie & Tieu, 2004; Zheng et al., 2013), has investigated RBF kernels with varying bandwidths in local regions. These studies consistently demonstrate the theoretical and simulated superiority of locally adaptive bandwidth estimators over global estimators. In the field of machine learning, successful experimental attempts have been achieved by directly applying asymmetric kernel functions (Moreno et al., 2003; Koide & Yamashita, 2006) or by incorporating them into existing kernel-based learning models along with asymmetric metric learning (Wu et al., 2010; Pintea et al., 2018).However, many of these studies lack a robust theoretical explanation, leaving the meaning of corresponding models and hypothesis space (no longer RKHS) still unknown. In this paper, for the first time, we demonstrate that it is actually an integral space of RKHSs.

Asymmetric kernel-based learning. Existing research in asymmetric kernel learning has primarily proposed frameworks based on SVD (Suykens, 2016) and least square SVM (He et al., 2023). However, for regression tasks, current works (Mackenzie & Tieu, 2004; Pintea et al., 2018) directly incorporate asymmetric kernels into symmetric-kernel-based learning models, lacking interpretability. Additionally, other works primarily focus on interpreting associated optimization models (Wu et al., 2010; Lin et al., 2022), where the corresponding functional space is regarded as a Reproducible Kernel Banach Space. This, however, is not currently applicable to LAB RBF kernels, as their reproducible property remains undetermined. Despite notable progress in theory, current applications of asymmetric kernel matrices often rely on datasets (e.g. the directed graph in (He et al., 2023)) or recognized asymmetric similarity measures (e.g. the Kullback-Leibler kernels in (Moreno et al., 2003)) This yields improved performance in specific scenarios but leaving a significant gap in addressing diverse datasets. With the help of trainable LAB RBF kernels, this paper proposes a robust groundwork for utilizing asymmetric kernels in tackling general regression tasks.

3. Asymmetric Kernel Ridge Regression

3.1. Kernel Ridge Regression

Kernel ridge regression (Vovk, 2013) is one of the most elementary kernelized algorithms. Define the dataset $\mathcal{X} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\} \subset \mathbb{R}^M, \mathcal{Y} = \{y_1, \cdots, y_N\} \subset \mathbb{R}$, and data matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N] \in \mathbb{R}^{M\times N}, \boldsymbol{Y} = [y_1, y_2, \cdots, y_N]^\top \in \mathbb{R}^N$. The task is to find a linear function in a high dimensional feature space, denoted as $\mathbb{R}^F$, which models the dependencies between the features $\phi(\boldsymbol{x}_i), \forall \boldsymbol{x}_i \in \mathcal{X}$ of input and response variables $y_i, \forall y_i \in \mathcal{Y}$. Here, $\phi : \mathbb{R}^M \to \mathbb{R}^F$ denotes the feature mapping from the data space to the feature space. Define $\phi(\boldsymbol{X}) = [\phi(\boldsymbol{x}_1), \phi(\boldsymbol{x}_2), \cdots, \phi(\boldsymbol{x}_N)]$, then the classical optimization model is as follow:

$$\min_{\boldsymbol{w}} \ \frac{\lambda}{2}\boldsymbol{w}^\top\boldsymbol{w} + \frac{1}{2}\|\boldsymbol{Y} - \phi(\boldsymbol{X})^\top\boldsymbol{w}\|_2^2, \tag{2}$$

where $\lambda > 0$ is a trade-off hyper-parameter. By utilizing the following well-known matrix inversion lemma (see (Petersen & Pedersen, 2008) for more information),

$$(\boldsymbol{A} + \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C})^{-1}\boldsymbol{B}\boldsymbol{D}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B} + \boldsymbol{D})^{-1}, \tag{3}$$

one can obtain the solution of KRR as follow

$$\begin{aligned}\boldsymbol{w}^* &= (\phi(\boldsymbol{X})\phi(\boldsymbol{X})^\top + \lambda\boldsymbol{I}_F)^{-1}\phi(\boldsymbol{X})\boldsymbol{Y} \\ &\overset{(a)}{=} \phi(\boldsymbol{X})(\lambda\boldsymbol{I}_N + \phi(\boldsymbol{X})^\top\phi(\boldsymbol{X}))^{-1}\boldsymbol{Y},\end{aligned}$$

where (3) is applied in (a) with $\boldsymbol{A} = \boldsymbol{I}_F, \ \boldsymbol{B} = \phi(\boldsymbol{X}), \ \boldsymbol{C} = \phi^\top(\boldsymbol{X}), \ \boldsymbol{D} = \boldsymbol{I}_N$.

3.2. Asymmetric Kernel Ridge Regression

Assume we have two feature mappings from data space to an unknown vector space: $\phi : \mathbb{R}^M \to \mathbb{R}^F$, and $\psi : \mathbb{R}^M \to \mathbb{R}^F$. Given training dataset $(\boldsymbol{X}, \boldsymbol{Y})$, the asymmetric kernel ridge regression model is

$$\begin{aligned}&\min_{\boldsymbol{w},\boldsymbol{v}} \ \lambda\boldsymbol{w}^\top\boldsymbol{v} + (\phi^\top(\boldsymbol{X})\boldsymbol{w} - \boldsymbol{Y})^\top(\psi^\top(\boldsymbol{X})\boldsymbol{v} - \boldsymbol{Y}) \\ &\Longleftrightarrow \min_{\boldsymbol{w},\boldsymbol{v}} \ \lambda\boldsymbol{w}^\top\boldsymbol{v} + \frac{1}{2}\|\phi^\top(\boldsymbol{X})\boldsymbol{w} - \boldsymbol{Y}\|_2^2 \\ &+ \frac{1}{2}\|\psi^\top(\boldsymbol{X})\boldsymbol{v} - \boldsymbol{Y}\|_2^2 - \frac{1}{2}\|\psi^\top(\boldsymbol{X})\boldsymbol{v} - \phi^\top(\boldsymbol{X})\boldsymbol{w}\|_2^2.\end{aligned} \tag{4}$$

Here, $\lambda > 0$ serves as a trade-off hyper-parameter between the regularization term $\boldsymbol{w}^\top\boldsymbol{v}$ and the error term