Kernel Subspace and Feature Extraction

Xiangxiang Xu and Lizhong Zheng
Dept. EECS, MIT
Cambridge, MA 02139, USA
Email: {xuxx, lizhong}@mit.edu

2023 IEEE International Symposium on Information Theory (ISIT) | 978-1-6654-7554-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISIT54713.2023.10206532

***Abstract*—We study kernel methods in machine learning from the perspective of feature subspace. We establish a one-to-one correspondence between feature subspaces and kernels and propose an information-theoretic measure for kernels. In particular, we construct a kernel from Hirschfeld–Gebelein–Rényi maximal correlation functions, coined the maximal correlation kernel, and demonstrate its information-theoretic optimality. We use the support vector machine (SVM) as an example to illustrate a connection between kernel methods and feature extraction approaches. We show that the kernel SVM on maximal correlation kernel achieves minimum prediction error. Finally, we interpret the Fisher kernel as a special maximal correlation kernel and establish its optimality.**

I. Introduction

One main objective of machine learning is to obtain useful information from often high-dimensional data. To this end, it is a common practice to extract meaningful feature representations from original data and then process features [1]. Neural networks [2] and kernel methods [3]–[6] are two of the most representative approaches to map data into feature space. In neural networks, the features are represented as the outputs of hidden neurons in the network. In contrast, the feature mapping in kernel methods is defined by the used kernel, which is used implicitly and is often infinite dimensional. While kernel approaches require much fewer parameters and can obtain good empirical performance on certain tasks [7], the performance significantly relies on the choice of kernels. With many attempts to investigate kernel methods [6], [8], [9], there still lacks a theoretical understanding of the mechanism behind kernel methods, which restricts their applications on complicated data.

On the other hand, the feature extraction in deep neural networks has been studied recently by information-theoretic and statistical analyses [10], [11]. For example, it was shown in [10] that, the feature extracted by deep neural networks coincides with the most informative feature, which is essentially related to the classical Hirschfeld–Gebelein–Rényi (HGR) maximal correlation problem [12]–[14]. Such theoretical characterizations provide a better understanding of existing algorithms and have been shown useful in designing algorithms for multimodal learning tasks [15].

In this paper, our goal is to characterize kernel methods from the perspective of feature subspace and reveal its connection with other learning approaches. We first introduce the associated kernel with each given feature subspace, which we coin the *projection kernel*, to establish a correspondence between kernel operations and geometric operations in feature subspaces. This connection allows us to study kernels methods via analyzing the corresponding feature subspaces. Specifically, we propose an information-theoretic measure for projection kernels, and demonstrate that the information-theoretically optimal kernel can be constructed from the HGR maximal correlation functions, coined the *maximal correlation kernel*. We further demonstrate that the support vector machine (SVM) with maximal correlation kernel can obtain the minimum prediction error, which justifies its optimality in learning tasks. Our analysis also reveals connections between SVM and other classification approaches including neural networks. Finally, we interpret the Fisher kernel, a classical kernel induced from parameterized distribution families [16], as a special case of maximal correlation kernels, thus demonstrating its optimality.

II. Preliminaries and Notations

Throughout this paper, we use X, Y to denote two random variables with alphabets $\mathcal{X}, \mathcal{Y}$, and denote their joint distribution and marginals as $P_{X,Y}$ and P_X, P_Y, respectively. We also use $\mathbb{E}[\cdot]$ to denote the expectation with respect to $P_{X,Y}$.

A. Feature Space

We adopt the notation convention introduced in [15], and let $\mathcal{F}_\mathcal{X} \triangleq \{\mathcal{X} \to \mathbb{R}\}$ denote the feature space formed by the (one-dimensional) features of X, with the geometry defined as follows. The inner product $\langle \cdot, \cdot \rangle_{\mathcal{F}_\mathcal{X}}$ on $\mathcal{F}_\mathcal{X}$ is defined as $\langle f_1, f_2 \rangle_{\mathcal{F}_\mathcal{X}} \triangleq \mathbb{E}_{P_X}[f_1(X) f_2(X)]$ for $f_1, f_2 \in \mathcal{F}_\mathcal{X}$. This induces a norm $\|\cdot\|_{\mathcal{F}_\mathcal{X}}$ with $\|f\|_{\mathcal{F}_\mathcal{X}} \triangleq \sqrt{\langle f, f \rangle_{\mathcal{F}_\mathcal{X}}}$ for $f \in \mathcal{F}_\mathcal{X}$. Then, for given $f \in \mathcal{F}_\mathcal{X}$ and subspace $\mathcal{G}$ of $\mathcal{F}_\mathcal{X}$, we denote the projection of f onto $\mathcal{G}$ as

$$\Pi(f; \mathcal{G}) \triangleq \arg\min_{h \in \mathcal{G}} \|h - f\|_{\mathcal{F}_\mathcal{X}}. \tag{1}$$

In addition, for a d-dimensional feature $f = (f_1, \ldots, f_d)^{\mathrm{T}} \colon \mathcal{X} \to \mathbb{R}^d$, we use $\operatorname{span}\{f\} \triangleq \operatorname{span}\{f_1, \ldots, f_d\}$ to denote the subspace spanned by all dimensions. We also use $\tilde{f}$ to denote the centered f, i.e., $\tilde{f}(x) \triangleq f(x) - \mathbb{E}_{P_X}[f(X)]$, and denote $\Lambda_f \triangleq \mathbb{E}_{P_X}\big[f(X) f^{\mathrm{T}}(X)\big]$.

B. Kernel

Given $\mathcal{X}$, $\mathcal{k} \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel on $\mathcal{X}$, if for all finite subset $\mathcal{I} \subset \mathcal{X}$, the $|\mathcal{I}|$ by $|\mathcal{I}|$ matrix $[\mathcal{k}(x, x')]_{x \in \mathcal{I}, x' \in \mathcal{I}}$ is positive semidefinite. For each kernel $\mathcal{k}$, we define the associated functional operator $\tau \colon \mathcal{F}_\mathcal{X} \to \mathcal{F}_\mathcal{X}$ as

$$[\tau(f)](x) \triangleq \mathbb{E}_{P_X}[\mathcal{k}(X, x) f(X)], \tag{2}$$

and we use $\mathsf{k} \leftrightarrow \tau$ to denote the correspondence between k and τ. Furthermore, we define the centered kernel $\tilde{\mathsf{k}}\colon \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ as

$$\tilde{\mathsf{k}}(x,x') \triangleq \mathsf{k}(x,x') - \bar{k}(x) - \bar{k}(x') + \mathbb{E}_{P_X}\big[\bar{k}(X)\big], \tag{3}$$

where we have defined $\bar{k}\colon (x \mapsto \mathbb{E}_{P_X}[\mathsf{k}(X,x)]) \in \mathcal{F}_{\mathcal{X}}$.

The following fact is the basis of the kernel trick in learning algorithms.

Fact 1: For each given kernel k, there exist an inner product space $\mathcal{V}$ with the inner product $\langle\cdot,\cdot\rangle_{\mathcal{V}}$, and a mapping $\nu\colon \mathcal{X} \to \mathcal{V}$, such that $\mathsf{k}(x,x') = \langle \nu(x), \nu(x')\rangle_{\mathcal{V}}$.

Remark 1: Suppose ν is one mapping for k satisfying Fact 1. Then for the centered kernel $\tilde{\mathsf{k}}$ [cf. (3)], we have $\tilde{\mathsf{k}}(x,x') = \langle \tilde{\nu}(x), \tilde{\nu}(x')\rangle_{\mathcal{V}}$, where $\tilde{\nu}(x) \triangleq \nu(x) - \mathbb{E}_{P_X}[\nu(X)]$.

In addition, we introduce the kernelized discriminative model (KDM) as follows.

Definition 1 (Kernelized Discriminative Model): For each kernel k, we define its associated kernelized discriminative model $P^{(\mathsf{k})}_{Y|X}$ as

$$P^{(\mathsf{k})}_{Y|X}(y|x) \triangleq P_Y(y)\Big(1 + \mathbb{E}\Big[\tilde{\mathsf{k}}(X,x)\Big|Y=y\Big]\Big). \tag{4}$$

Then, we use $\hat{y}^{(\mathsf{k})}$ to denote the maximum a posteriori (MAP) estimation induced from KDM $P^{(\mathsf{k})}_{Y|X}$, i.e.,

$$\hat{y}^{(\mathsf{k})}(x) \triangleq \arg\max_{y\in\mathcal{Y}} P^{(\mathsf{k})}_{Y|X}(y|x). \tag{5}$$

The KDM can be regarded as a generalized probability distribution, since we have $\sum_{y\in\mathcal{Y}} P^{(\mathsf{k})}_{Y|X}(y|x) = 1$ for all $x \in \mathcal{X}$ while $P^{(\mathsf{k})}_{Y|X}(y|x)$ can sometimes take negative values.

C. Modal Decomposition, Maximal Correlation, and H-score

We first introduce the modal decomposition of joint distribution $P_{X,Y}$ [11], [15].

Proposition 1 (Modal Decomposition [11]): For given $P_{X,Y}$, there exists $K \le \min\{|\mathcal{X}|, |\mathcal{Y}|\} - 1$, such that

$$P_{X,Y}(x,y) = P_X(x)P_Y(y)\left(1 + \sum_{i=1}^{K} \sigma_i f_i^*(x) g_i^*(y)\right), \tag{6}$$

where $\sigma_1 \ge \sigma_2 \ge \sigma_K > 0$, and $\mathbb{E}\big[f_i^*(X) f_j^*(X)\big] = \mathbb{E}\big[g_i^*(Y) g_j^*(Y)\big] = \mathbb{1}_{\{i=j\}}$ for all $1 \le i,j \le K$, where $\mathbb{1}_{\{\cdot\}}$ denotes the indicator function.

It can be shown that (f_i^*, g_i^*) pairs are the most correlated function pairs of X and Y, referred to as maximal correlation functions. We also denote $\varrho \triangleq \sigma_1$, known as the HGR maximal correlation [12]–[14] of X and Y, and define the K-dimensional feature $f^*(x) \triangleq [f_1^*(x), \dots, f_K^*(x)]^{\mathrm{T}}$. In particular, when Y is binary, we have $f^* = f_1^* \in \mathcal{F}_{\mathcal{X}}$.

It has been shown in [11] that the maximal correlation functions $f_i^*, i = 1, \dots, K$ are the optimal features of X in inferring or estimating Y. In general, given a d-dimensional feature f of X, the effectiveness of f in inferring or estimating Y can be measured by its *H-score* [10], [11], defined as

$$\mathscr{H}(f) \triangleq \frac{1}{2}\cdot \mathbb{E}\left[\left\|\mathbb{E}\Big[\Lambda_f^{-\frac{1}{2}} \tilde{f}(X)\Big|Y\Big]\right\|^2\right], \tag{7}$$

where $\tilde{f}(x) \triangleq f(x) - \mathbb{E}[f(X)]$. It can be verified that for all d and $f\colon \mathcal{X} \to \mathbb{R}^d$, we have

$$\mathscr{H}(f) \le \mathscr{H}(f^*) = \frac{1}{2}\sum_{i=1}^{K} \sigma_i^2, \tag{8}$$

where $\sigma_1, \dots, \sigma_K$ are as defined in (6).

D. Binary Classification

We consider the binary classification problem which predicts binary label Y from the data variable X. For convenience, we assume Y takes values from $\mathcal{Y} \triangleq \{-1, 1\}$.

Suppose the training dataset contains n sample pairs $\{(x_i, y_i)\}_{i=1}^n$ of (X,Y), and let $P_{X,Y}$ denote the corresponding empirical distribution, i.e.,

$$P_{X,Y}(x,y) \triangleq \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{x_i=x, y_i=y\}}. \tag{9}$$

1) Support Vector Machine: The support vector machine (SVM) solves binary classification tasks by finding the optimal hyperplane that separates two classes with maximum margin [7]. Given d-dimensional feature mapping $f\colon \mathcal{X} \to \mathbb{R}^d$, the loss for SVM based on f can be written as

$$\begin{aligned} &L_{\mathsf{SVM}}(f,w,b;\lambda) \\ &\quad\triangleq \mathbb{E}_{P_{X,Y}}[\ell_{\mathrm{hinge}}(Y, \langle w, f(X)\rangle + b)] + \frac{\lambda}{2}\cdot\|w\|^2, \end{aligned} \tag{10}$$

where $w, b \in \mathbb{R}^d$ are the parameters of the hyperplane, where $\lambda > 0$ is a hyperparameter of SVM, and where $\ell_{\mathrm{hinge}}\colon \mathcal{Y}\times\mathbb{R} \to \mathbb{R}$ denotes the hinge loss, defined as $\ell_{\mathrm{hinge}}(y,z) \triangleq (1-yz)^+$ with $x^+ \triangleq \max\{0,x\}$.

Moreover, let $(w_{\mathsf{SVM}}, b_{\mathsf{SVM}}) \triangleq \arg\min_{w,b} L_{\mathsf{SVM}}(f,w,b;\lambda)$ and $L^*_{\mathsf{SVM}}(f;\lambda) \triangleq L_{\mathsf{SVM}}(f, w_{\mathsf{SVM}}, b_{\mathsf{SVM}};\lambda)$ denote the optimal parameters and the value of loss function, respectively. Then, the prediction of SVM is

$$\hat{y}_{\mathsf{SVM}}(x; f, \lambda) \triangleq \mathrm{sgn}(\langle w_{\mathsf{SVM}}, f(x)\rangle + b_{\mathsf{SVM}}), \tag{11}$$

where $\mathrm{sgn}(\cdot)$ denotes the sign function.

Specifically, for a given kernel k, the prediction of the corresponding *kernel SVM* is[1] $\hat{y}^{(\mathsf{k})}_{\mathsf{SVM}}(x;\lambda) \triangleq \hat{y}_{\mathsf{SVM}}(x;\nu,\lambda)$, where ν is any mapping given by Fact 1.

2) Logistic Regression and Neural Networks: Given d-dimensional feature f of X, the discriminative model of logistic regression is $\tilde{P}_{Y|X}(y|x;f,w,b) \triangleq \mathrm{sigmoid}(y\cdot(\langle w, f(x)\rangle + b))$, where $w \in \mathbb{R}^d, b \in \mathbb{R}$ are the weight and bias, respectively, and where $\mathrm{sigmoid}(\cdot)$ is defined as $\mathrm{sigmoid}(x) \triangleq \frac{1}{1+\exp(-x)}$.

Then, the loss of logistic regression is $L_{\mathsf{LR}}(f,w,b) \triangleq -\mathbb{E}\Big[\log \tilde{P}_{Y|X}(Y|X;f,w,b)\Big]$, and the optimal parameters $w_{\mathsf{LR}}, b_{\mathsf{LR}}$ are learned by minimizing the loss, i.e., $(w_{\mathsf{LR}}, b_{\mathsf{LR}}) \triangleq \arg\min_{w,b} L_{\mathsf{LR}}(f,w,b)$. The resulting decision rule is

$$\hat{y}_{\mathsf{LR}}(x;f) \triangleq \arg\max_{y\in\mathcal{Y}} \tilde{P}_{Y|X}(y|x;f,w_{\mathsf{LR}},b_{\mathsf{LR}})$$

[1]It is worth mentioning that the practical implementation of kernel SVM is typically done by solving a dual optimization problem without explicitly using ν. See [17, Section 12] for detailed discussions.

$$= \operatorname{sgn}(\langle w_{\mathsf{LR}}, f(x)\rangle + b_{\mathsf{LR}}). \tag{12}$$

The logistic regression is often used as the classification layer for multi-layer neural networks, where w and b correspond to weights and the bias term, respectively. In this case, the feature mapping $f(\cdot)$ also takes a parameterized form, and the parameters of $f(\cdot)$ are jointly learned with w and b.

Due to the space limitations, we omit most proofs in the rest of this paper, but refer the readers to the extended version of this paper [18] for the details.

III. Projection Kernel and Informative Features

In this section, we introduce a one-to-one correspondence between kernels and feature subspaces, and then characterize the informativeness of kernels by investigating the features in the associated subspaces.

A. Projection Kernel and Feature Subspace

We first introduce a family of kernels with one-to-one correspondence to feature subspace.

Definition 2 (Projection Kernel): Let $\mathcal{G}$ denote a d-dimensional subspace of $\mathcal{F}_{\mathcal{X}}$ with a basis $\{f_1, \ldots, f_d\}$. We use $\mathcal{k}_{\mathcal{G}} \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ to denote the projection kernel associated with $\mathcal{G}$, defined as $\mathcal{k}_{\mathcal{G}}(x, x') \triangleq f^{\mathrm{T}}(x)\Lambda_f^{-1} f(x')$, where we have defined $f \triangleq (f_1, \ldots, f_d)^{\mathrm{T}}$ and $\Lambda_f \triangleq \mathbb{E}\big[f(X)f^{\mathrm{T}}(X)\big]$.

With slight abuse of notation, we also dnenote $\mathcal{k}_f \triangleq \mathcal{k}_{\operatorname{span}\{f\}}$, the projection kernel associated with $\operatorname{span}\{f\}$.

Note that $\mathcal{k}_{\mathcal{G}}$ is a valid kernel function, and the corresponding ν mapping in Fact 1 can be chosen as $\nu(x) = [f_1(x), \ldots, f_d(x)]^{\mathrm{T}}$ for any orthonormal basis $\{f_1, \ldots, f_d\}$ of $\mathcal{G}$. It turns out that the functional operators associated with projection kernels are projection operators in the feature space, which we formalize as follows.

Property 1: Let $\tau \leftrightarrow \mathcal{k}_{\mathcal{G}}$ denote the operator corresponding to subspace $\mathcal{G}$ [cf. (2)], then we have $\tau(f) = \Pi(f; \mathcal{G})$ for all $f \in \mathcal{F}_{\mathcal{X}}$.

Therefore, given a projection kernel $\mathcal{k}$, the associated subspace can be represented as $\{f \in \mathcal{F}_{\mathcal{X}} \colon \tau(f) = f\}$, where $\tau \leftrightarrow \mathcal{k}$ is the associated operator. This establishes a one-to-one correspondence between projection kernels and feature subspaces.

B. H-score and Informative Features

The projection kernel provides a connection between feature subspace and kernel, from which we can characterize subspace $\mathcal{G}$ in terms of the corresponding kernel $\mathcal{k}_{\mathcal{G}}$. Specifically, we can represent the H-score [cf. (7)] of a feature f in terms of the projection kernel $\mathcal{k}_f$, formalized as follows.

Proposition 2: For all f with $\operatorname{span}\{f\} = \mathcal{G}$, we have $\mathscr{H}(f) = \frac{1}{2} \cdot \big(\mathbb{E}_{P_{XX'}}[\mathcal{k}_{\mathcal{G}}(X, X')] - \mathbb{E}_{P_X P_{X'}}[\mathcal{k}_{\mathcal{G}}(X, X')]\big)$, where we have defined X' such that the joint distribution of X and X' is

$$P_{XX'}(x, x') \triangleq \sum_{y \in \mathcal{Y}} P_Y(y) P_{X|Y=y}(x) P_{X|Y=y}(x'). \tag{13}$$

With slight abuse of notation, we can use $\mathscr{H}(\mathcal{G})$ to denote the H-score corresponding to feature subspace $\mathcal{G}$. In particular, we have the following characterization of $\mathscr{H}(\mathcal{G})$ when Y is binary.

Proposition 3: Suppose Y is binary, and f^* is the maximal correlation function of $P_{X,Y}$. Then, for each subspace $\mathcal{G}$ of $\mathcal{F}_{\mathcal{X}}$, we have

$$\mathscr{H}(\mathcal{G}) = \frac{\varrho^2}{2} \cdot \big\|\Pi(f^*; \mathcal{G})\big\|_{\mathcal{F}_{\mathcal{X}}}^2 = \max_{f \in \mathcal{G}} \mathscr{H}(f) = \mathscr{H}(\Pi(f^*; \mathcal{G})). \tag{14}$$

From Proposition 3, $\mathscr{H}(\mathcal{G})$ depends only on the projection of f^* onto $\mathcal{G}$, which is also the most informative feature in $\mathcal{G}$. In addition, note that since $\|f^*\|_{\mathcal{F}_{\mathcal{X}}} = 1$, $\big\|\Pi(f^*; \mathcal{G})\big\|_{\mathcal{F}_{\mathcal{X}}}^2$ is also the cosine value of the principal angle between f^* and $\mathcal{G}$. Therefore, we can interpret the H-score as a measure of the principal angle between the optimal feature f^* and the given subspace.

C. Maximal Correlation Kernel

Note that from (8), $\mathscr{H}(f)$ is maximized when f takes the maximal correlation function f^*. Therefore, the subspace $\operatorname{span}\{f^*\}$ (and thus projection kernel $\mathcal{k}_{f^*}$) is optimal in terms of the H-score measure. We will denote $\mathcal{k}^* \triangleq \mathcal{k}_{f^*}$, referred to as the *maximal correlation kernel*.

Specifically, the KDM (cf. Definition 1) of maximal correlation kernel $\mathcal{k}^*$ coincides with the underlying conditional distribution $P_{Y|X}$, demonstrated as follows.

Property 2: For all x and y, we have $P_{Y|X}(y|x) = P_{Y|X}^{(\mathcal{k}^*)}(y|x)$ and $\hat{y}^{(\mathcal{k}^*)}(x) = \hat{y}_{\mathsf{MAP}}(x)$, where $\hat{y}_{\mathsf{MAP}}$ denotes the MAP estimation, i.e.,

$$\hat{y}_{\mathsf{MAP}}(x) \triangleq \arg\max_{y \in \mathcal{Y}} P_{Y|X}(y|x). \tag{15}$$

As we will develop in the next section, the maximal correlation kernel also achieves the optimal performance in support vector machine.

IV. Support Vector Machine Analysis

In this section, we investigate support vector machine, a representative kernel approach for binary classification. Let (X, Y) denote the training data and corresponding label taken from $\mathcal{Y} = \{-1, 1\}$, with $P_{X,Y}$ denoting the empirical distribution as defined in (9). Throughout this section, we will focus on the balanced dataset with

$$P_Y(-1) = P_Y(1) = \frac{1}{2}. \tag{16}$$

It can be verified that in this case, the MAP estimation [cf. (15)] can be expressed in terms of maximal correlation function.

Property 3: Under assumption (16), we can express the MAP estimation as $\hat{y}_{\mathsf{MAP}}(x) = \operatorname{sgn}(f^*(x))$ for all $x \in \mathcal{X}$, where $f^* \in \mathcal{F}_{\mathcal{X}}$ is the maximal correlation function of $P_{X,Y}$.

A. SVM on Given Features

We first consider the SVM algorithm applied on a given feature representation $f(X) \in \mathbb{R}^d$, which can also be regarded as the kernel SVM on kernel $\mathcal{k}(x, x') = \langle f(x), f(x')\rangle$.

To begin, for each given feature f and $\lambda > 0$, let us define

$$\hat{L}(f;\lambda) \triangleq 1 - \frac{1}{2\lambda} \cdot \|\mathbb{E}[f(X)Y]\|^2.$$

Then we have the following characterization.

Theorem 1: For all given feature f and $\lambda \geq 0$, we have

$$\hat{L}(f;\lambda) \leq L^*_{\mathsf{SVM}}(f;\lambda) \leq \hat{L}(f;\lambda) + \left(\frac{\lambda_{\mathrm{T}}}{\lambda} - 1\right)^+, \quad (17)$$

where we have defined $\lambda_{\mathrm{T}} \triangleq M \cdot \|\mathbb{E}[f(X)Y]\|$ and $M \triangleq \max_{x\in\mathcal{X}} \big\|\tilde{f}(x)\big\|$, with $\tilde{f}(x) \triangleq f(x) - \mathbb{E}[f(X)]$, and where $x^+ \triangleq \max\{0, x\}$.

Specifically, when $\lambda \geq \lambda_{\mathrm{T}}$, we have $L^*_{\mathsf{SVM}}(f;\lambda) = \hat{L}(f;\lambda)$, which can be achieved by

$$w_{\mathsf{SVM}} = \frac{1}{\lambda} \cdot \mathbb{E}[f(X)Y], \quad b_{\mathsf{SVM}} = -\langle w_{\mathsf{SVM}}, \mathbb{E}[f(X)]\rangle, \quad (18)$$

and the resulting SVM prediction is

$$\hat{y}_{\mathsf{SVM}}(x; f, \lambda) = \mathrm{sgn}\Big(\Big\langle \mathbb{E}\big[\tilde{f}(X)Y\big], \tilde{f}(x)\Big\rangle\Big) \quad (19)$$

$$= \arg\min_{y\in\mathcal{Y}} \|f(x) - \mathbb{E}[f(X)|Y = y]\|. \quad (20)$$

From Theorem 1, when $\lambda \geq \lambda_{\mathrm{T}}$, the SVM decision $\hat{y}_{\mathsf{SVM}}(x; f, \lambda)$ does not depend on the value of λ. In the remaining, we will focus on the regime where $\lambda \geq \lambda_{\mathrm{T}}$, and drop the λ in expressions whenever possible, e.g., we simply denote $\hat{y}_{\mathsf{SVM}}(x; f, \lambda)$ by $\hat{y}_{\mathsf{SVM}}(x; f)$. As we will see soon, SVM can still obtain minimum prediction error in this regime, by using a good feature mapping f (or equivalently, a good kernel).

From (20), the SVM prediction can be interpreted as a nearest centroid classifier, where decision is based on comparing the distance between $f(x)$ and the class centroids $\mathbb{E}[f(X)|Y = y]$, $y \in \mathcal{Y}$. In addition, from

$$\mathbb{E}[f(X)Y] = \mathbb{E}[Y \cdot \mathbb{E}[f(X)|Y]]$$
$$= \frac{1}{2}(\mathbb{E}[f(X)|Y = 1] - \mathbb{E}[f(X)|Y = -1]),$$

we can interpret the SVM loss $L^*_{\mathsf{SVM}} = \hat{L}$ as measuring the distance between two class centroids.

Furthermore, when f is one-dimensional feature, we can rewrite (19) as

$$\hat{y}_{\mathsf{SVM}}(x; f) = \mathrm{sgn}\Big(\Big\langle \mathbb{E}\big[\tilde{f}(X)Y\big], \tilde{f}(x)\Big\rangle\Big) = \mathrm{sgn}\Big(\hat{f}(x)\Big),$$

where $\hat{f} \triangleq \Pi\Big(f^*; \mathrm{span}\{\tilde{f}\}\Big)$. Therefore, the decision rule depends only the projection of f^* onto the subspace $\mathrm{span}\{\hat{f}\}$, which is also the most informative features on the subspace (cf. Proposition 3). Later on we will see a similar geometric illustration of kernel SVM.

Moreover, we can establish a connection between SVM loss and the H-score measure, formalized as the following corollary.

Corollary 1: Suppose $\lambda \geq \lambda_{\mathrm{T}}$, then we have

$$1 - \frac{r_{\max}}{\lambda} \cdot \mathscr{H}(\tilde{f}) \leq L^*_{\mathsf{SVM}}(f;\lambda) \leq 1 - \frac{r_{\min}}{\lambda} \cdot \mathscr{H}(\tilde{f}),$$

where $r_{\max}$ and $r_{\min}$ denote the maximum and minimum positive eigenvalues of the covariance matrix $\Lambda_{\tilde{f}}$, respectively. Specifically, if $\Lambda_{\tilde{f}} = I$, then we have $L^*_{\mathsf{SVM}}(f;\lambda) = 1 - \lambda^{-1} \cdot \mathscr{H}(\tilde{f})$.

As a result, for each normalized feature f with covariance matrix $\Lambda_{\tilde{f}} = I$, the SVM loss L^*_{SVM} measures the informativeness of f in inferring the label Y.

B. Kernel SVM

In practice, instead of applying SVM on a given or manually designed feature f, it is more often to directly implement SVM on a kernel $\mathcal{k}$. Similar to Theorem 1, we have the following characterization, from which we can interpret KDM as a probabilistic output for kernel SVM.

Theorem 2: For each given kernel $\mathcal{k}$, there exists a constant $\lambda_{\mathrm{T}} > 0$, such that when $\lambda \geq \lambda_{\mathrm{T}}$, the SVM prediction is $\hat{y}^{(\mathcal{k})}_{\mathsf{SVM}}(x) = \mathrm{sgn}([\tau(f^*)](x))$, where $\tau \leftrightarrow \tilde{\mathcal{k}}$ is the operator associated with centered kernel $\tilde{\mathcal{k}}$ [cf. (2) and (3)]. In addition, the SVM prediction coincides with the KDM prediction (cf. Definition 1) obtained from $\mathcal{k}$, i.e., we have $\hat{y}^{(\mathcal{k})}_{\mathsf{SVM}}(x) = \hat{y}^{(\mathcal{k})}(x)$ for all $x \in \mathcal{X}$.

Proof: Let $\mathcal{V}$ and $\nu\colon \mathcal{X} \to \mathcal{V}$ denote the inner product space and mapping associated with kernel $\mathcal{k}$ (cf. Fact 1), and let $\tilde{\nu}(x) \triangleq \nu(x) - \mathbb{E}_{P_X}[\nu(X)]$. Then, we have

$$\langle \mathbb{E}[\tilde{\nu}(X)Y], \tilde{\nu}(x)\rangle_{\mathcal{V}} = \mathbb{E}[\langle \tilde{\nu}(X), \tilde{\nu}(x)\rangle_{\mathcal{V}} \cdot Y]$$
$$= \mathbb{E}\Big[\tilde{\mathcal{k}}(X, x) \cdot Y\Big], \quad (21)$$

which can be rewritten as

$$\mathbb{E}\Big[\tilde{\mathcal{k}}(X, x) \cdot Y\Big]$$
$$= \mathbb{E}_{P_{X,Y}}\Big[\tilde{\mathcal{k}}(X, x) \cdot Y\Big]$$
$$= \mathbb{E}_{P_X P_Y}\Big[\tilde{\mathcal{k}}(X, x) \cdot Y \cdot (1 + \varrho \cdot f^*(X) \cdot Y)\Big]$$
$$= \mathbb{E}\Big[\tilde{\mathcal{k}}(X, x)\Big] \cdot \mathbb{E}[Y] + \varrho \cdot \mathbb{E}\Big[\tilde{\mathcal{k}}(X, x) f^*(X)\Big] \cdot \mathbb{E}\big[Y^2\big]$$
$$= \varrho \cdot \mathbb{E}\Big[\tilde{\mathcal{k}}(X, x) f^*(X)\Big]$$
$$= \varrho \cdot [\tau(f^*)](x),$$

where to obtain the second equality we have used the modal decomposition of $P_{X,Y}$.

Hence, from Theorem 1 we obtain

$$\hat{y}^{(\mathcal{k})}_{\mathsf{SVM}}(x) = \hat{y}_{\mathsf{SVM}}(x; \nu) = \mathrm{sgn}(\langle \mathbb{E}[\tilde{\nu}(X)Y], \tilde{\nu}(x)\rangle)$$
$$= \mathrm{sgn}\Big(\mathbb{E}\Big[\tilde{\mathcal{k}}(X, x) \cdot Y\Big]\Big)$$
$$= \mathrm{sgn}([\tau(f^*)](x)).$$

It remains only to establish the equivalence between $\hat{y}^{(\mathcal{k})}_{\mathsf{SVM}}$ and KDM decision $\hat{y}^{(\mathcal{k})}$. To this end, note that from (4) and the balanced dataset assumption (16), we have

$$P^{(\mathcal{k})}_{Y|X}(y|x) = P_Y(y)\Big(1 + \mathbb{E}\Big[\tilde{\mathcal{k}}(X, x)\Big|Y = y\Big]\Big)$$
$$= \frac{1}{2}\Big(1 + y \cdot \mathbb{E}\Big[\tilde{\mathcal{k}}(X, x)Y\Big]\Big)$$

for all $x \in \mathcal{X}, y \in \mathcal{Y}$.

Hence, for all $x \in \mathcal{X}$,

$$\hat{y}^{(\mathcal{k})}(x) = \arg\max_{y \in \mathcal{Y}} P^{(\mathcal{k})}_{Y|X}(y|x) = \operatorname{sgn}\Big(\mathbb{E}\Big[\tilde{\mathcal{k}}(X,x)Y\Big]\Big) = \hat{y}^{(\mathcal{k})}_{\mathsf{SVM}}(x),$$

which completes the proof. ∎

From Theorem 2, the final decision $\hat{y}^{(\mathcal{k})}_{\mathsf{SVM}}$ depends on $\mathcal{k}$ only through the centered kernel $\tilde{\mathcal{k}}$. Moreover, compare Theorem 2 with Property 3, kernel SVM prediction differs from MAP only in applying the operator τ on f^*. In particular, when the maximal correlation function f^* is an eigenfunction of the corresponding operator $\tau \leftrightarrow \tilde{\mathcal{k}}$, i.e., $\tau(f^*) = c \cdot f^*$ for some $c > 0$, the SVM prediction coincides with the MAP prediction, i.e., $\hat{y}^{(\mathcal{k})}_{\mathsf{SVM}}(x) = \hat{y}_{\mathsf{MAP}}(x)$ for all $x \in \mathcal{X}$.

If we restrict our attention to projection kernels, the kernel SVM decision can be further interpreted as a projection operation on the associated subspace. To see this, let $\mathcal{G}$ denote a feature subspace of $\mathcal{F}_{\mathcal{X}}$ spanned by zero-mean features, then from Theorem 1 and Proposition 3, the kernel SVM loss for $\mathcal{k}_{\mathcal{G}}$ is

$$1 - \frac{1}{\lambda} \cdot \mathscr{H}(\mathcal{G}) = 1 - \frac{\varrho^2}{2\lambda} \cdot \big\|\Pi(f^*; \mathcal{G})\big\|^2_{\mathcal{F}_{\mathcal{X}}},$$

which measures the principal angle between f^* and $\mathcal{G}$. In addition, the decision rule can be expressed as

$$\hat{y}^{(\mathcal{k}_{\mathcal{G}})}_{\mathsf{SVM}}(x) = \operatorname{sgn}([\Pi(f^*; \mathcal{G})](x)), \tag{22}$$

From Proposition 3, $\Pi(f^*; \mathcal{G})$ is also the most informative feature in $\mathcal{G}$. Therefore, kernel SVM on $\mathcal{k}_{\mathcal{G}}$ is equivalent to first extracting the most informative feature in $\mathcal{G}$, and then using the extracted feature to make decision.

C. Relationship to Other Classification Approaches

1) Maximum a Posteriori (MAP) Estimation: From (22), when the maximal correlation kernel $\mathcal{k}^*$ is applied, the kernel SVM decision is $\operatorname{sgn}(f^*(x))$, which coincides with the MAP prediction (cf. Property 3). Since MAP achieves the minimum prediction error, kernel SVM on the maximal correlation kernel also obtains the minimum prediction error.

2) Logistic Regression and Neural Networks: We have interpreted SVM as extracting the most informative feature, where the informativeness is measured by H-score. The analysis in [10] has shown that logistic regression is also equivalent to maximizing the H-score, when X and Y are weakly independent. Indeed, we can show that SVM and logistic regression lead to the same prediction in a weak dependence regime, which we formalize as follows.

Proposition 4: Suppose $\varrho = O(\epsilon)$ for some $\epsilon > 0$. For SVM and logistic regression applied on feature $f\colon \mathcal{X} \to \mathbb{R}^d$ with covariance $\Lambda_f = I_d$, the optimal parameters satisfy

$$w_{\mathsf{LR}} = 2\lambda \cdot w_{\mathsf{SVM}} + o(\epsilon),$$
$$b_{\mathsf{LR}} = 2\lambda \cdot b_{\mathsf{SVM}} + o(\epsilon),$$

where λ is the hyperparameter in SVM. In addition, we have $\hat{y}_{\mathsf{SVM}}(x; f) = \hat{y}_{\mathsf{LR}}(x; f)$ for ϵ sufficiently small.

Remark 2: Since H-score can also be directly maximized by implementing the maximal correlation regression [19], a similar connection holds for SVM and maximal correlation regression.

V. Fisher Kernel

We demonstrate that *Fisher kernel* [16], [20] can also be interpreted as a maximal correlation kernel.

Given a family of distributions $\pi(\cdot; \theta)$ supported on $\mathcal{X}$ and parameterized by $\theta \in \mathbb{R}^m$, suppose the score function $s_\theta(x) \triangleq \frac{\partial}{\partial \theta} \log \pi(x; \theta)$ exists. Then, the Fisher kernel is defined as the projection kernel associated with the score function s_θ, i.e., $\mathcal{k}_{s_\theta}$.

Specifically, we consider classification tasks where the joint distribution between data variable X and label Y are a mixture of the parameterized forms. Suppose for each class $Y = y \in \mathcal{Y}$, the data variable X is generated from

$$P_{X|Y}(x|y) = \pi(x; \theta_y) \tag{23}$$

for some $\theta_y \in \mathbb{R}^m$. Then we have the following result.

Theorem 3: Suppose $\|\theta_y\| < \epsilon$ for all $y \in \mathcal{Y}$, and let $s(x) \triangleq s_0(x)$. Then for the joint distribution $P_{X,Y} = P_{X|Y}P_Y$ generated according to (23), we have

$$P_{X,Y}(x,y) = P_X(x)P_Y(y)\Big(1 + \langle s(x), \tilde{\theta}_y \rangle\Big) + o(\epsilon), \tag{24}$$

$$\mathcal{k}_s(x,x') = \mathcal{k}^*(x,x') + o(\epsilon), \tag{25}$$

where $\tilde{\theta}_y \triangleq \theta_y - \mathbb{E}[\theta_Y]$ denotes the centered θ_y, and where $\mathcal{k}^*$ is the maximal correlation kernel defined on $P_{X,Y}$. In addition, the H-score of s satisfies

$$\mathscr{H}(s) = I(X;Y) + o(\epsilon^2), \tag{26}$$

where $I(X;Y)$ denotes the mutual information between X and Y.

From (24), the score function s is equal to the maximal correlation function f^* of $P_{X,Y}$ up to a linear transformation [cf. (6)], and we have $P_{Y|X}(y|x) = P^{(\mathcal{k}^*)}_{Y|X}(y|x) = P^{(\mathcal{k}_s)}_{Y|X}(y|x) + o(\epsilon)$. Therefore, Fisher kernel is the optimal kernel for tasks generated from (23).

VI. Conclusion

In this paper, we study kernel methods from the perspective of feature subspace, where we demonstrate a connection between kernel methods and informative feature extraction problems. With SVM as an example, we illustrate the relationship between kernel methods and neural networks. The theoretical results can help guide practical kernel designs and incorporate kernel methods with feature-based learning approaches.

Acknowledgments

This work was supported in part by the National Science Foundation (NSF) under Award CNS-2002908 and the Office of Naval Research (ONR) under grant N00014-19-1-2621.

REFERENCES

[1] D. Storcheus, A. Rostamizadeh, and S. Kumar, "A survey of modern questions and challenges in feature extraction," in *Feature Extraction: Modern Questions and Challenges*. PMLR, 2015, pp. 1–18.

[2] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[3] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K.-R. Mullers, "Fisher discriminant analysis with kernels," in *Neural networks for signal processing IX: Proceedings of the 1999 IEEE signal processing society workshop (cat. no. 98th8468)*. Ieee, 1999, pp. 41–48.

[4] F. R. Bach and M. I. Jordan, "Kernel independent component analysis," *Journal of machine learning research*, vol. 3, no. Jul, pp. 1–48, 2002.

[5] T. Hofmann, B. Schölkopf, and A. J. Smola, "Kernel methods in machine learning," *The annals of statistics*, vol. 36, no. 3, pp. 1171–1220, 2008.

[6] B. Schölkopf, "The kernel trick for distances," *Advances in neural information processing systems*, vol. 13, 2000.

[7] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[8] J. Platt *et al.*, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in large margin classifiers*, vol. 10, no. 3, pp. 61–74, 1999.

[9] J. Xu, R. Jenssen, A. Paiva, and I. Park, "A reproducing kernel hilbert space framework for itl," in *Information Theoretic Learning*. Springer, 2010, pp. 351–384.

[10] X. Xu, S.-L. Huang, L. Zheng, and G. W. Wornell, "An information theoretic interpretation to deep neural networks," *Entropy*, vol. 24, no. 1, p. 135, 2022.

[11] S.-L. Huang, A. Makur, G. W. Wornell, and L. Zheng, "On universal features for high-dimensional learning and inference," *arXiv preprint arXiv:1911.09105*, 2019.

[12] H. O. Hirschfeld, "A connection between correlation and contingency," in *Proceedings of the Cambridge Philosophical Society*, vol. 31, no. 4, 1935, pp. 520–524.

[13] H. Gebelein, "Das statistische problem der korrelation als variations-und eigenwertproblem und sein zusammenhang mit der ausgleichsrechnung," *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik*, vol. 21, no. 6, pp. 364–379, 1941.

[14] A. Rényi, "On measures of dependence," *Acta Mathematica Academiae Scientiarum Hungarica*, vol. 10, no. 3-4, pp. 441–451, 1959.

[15] X. Xu and L. Zheng, "Multivariate feature extraction," in *2022 58th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2022, pp. 1–8.

[16] T. S. Jaakkola, D. Haussler *et al.*, "Exploiting generative models in discriminative classifiers," *Advances in neural information processing systems*, pp. 487–493, 1999.

[17] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2009, vol. 2.

[18] X. Xu and L. Zheng, "Kernel subspace and feature extraction," *arXiv preprint arXiv:2301.01410*, 2023.

[19] X. Xu and S.-L. Huang, "Maximal correlation regression," *IEEE Access*, vol. 8, pp. 26 591–26 601, 2020.

[20] J. Shawe-Taylor, N. Cristianini *et al.*, *Kernel methods for pattern analysis*. Cambridge university press, 2004.