Feature Elimination in Empirical Risk Minimization and Support Vector Machines

Sayan Dasgupta∗ Yair Goldberg†
Michael R. Kosorok‡

∗University of North Carolina at Chapel Hill, sdg@live.unc.edu
†University of Haifa, ygoldberg@stat.haifa.ac.il
‡University of North Carolina at Chapel Hill, kosorok@unc.edu
This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.
http://biostats.bepress.com/uncbiostat/art37
Copyright ©2013 by the authors.
Feature Elimination in Empirical Risk Minimization and Support Vector Machines

Sayan Dasgupta, Yair Goldberg, and Michael R. Kosorok

Abstract

We develop an approach for feature elimination in empirical risk minimization and support vector machines, based on recursive elimination of features. We present theoretical properties of this method and show that this is uniformly consistent in finding the correct feature space under certain generalized assumptions. We present case studies to show that the assumptions are met in most practical situations and also present simulation studies to demonstrate performance of the proposed approach.
FEATURE ELIMINATION IN EMPIRICAL RISK MINIMIZATION AND SUPPORT VECTOR MACHINES

BY SAYAN DASGUPTA\textsuperscript{1}, YAIR GOLDBERG\textsuperscript{2} AND MICHAEL R. KOSOROK\textsuperscript{1}

\textsuperscript{1}The University of North Carolina at Chapel Hill
\textsuperscript{2}University of Haifa

We develop an approach for feature elimination in empirical risk minimization and support vector machines, based on recursive elimination of features. We present theoretical properties of this method and show that this is uniformly consistent in finding the correct feature space under certain generalized assumptions. We present case studies to show that the assumptions are met in most practical situations and also present simulation studies to demonstrate performance of the proposed approach.

1. Introduction. In recent years it has become increasingly easy to collect large amount of information, especially with respect to the number of explanatory variables or ‘features’. However the additional information provided by each of these features may not be significant for explaining the phenomenon at hand. Learning the functional connection between the explanatory variables and the response from such high-dimensional data can itself be quite challenging. Moreover some of these explanatory variables or features may contain redundant or noisy information and this may hamper the quality of learning. One way to overcome this problem is to use variable selection (also referred to as feature elimination) techniques to find a smaller set of variables that is able to perform the learning task sufficiently well.

In this work we discuss feature elimination in empirical risk minimization and support vector machines, focusing mainly on the latter. The popularity of support vector machines (SVM) as a set of supervised learning algorithms is motivated by the fact that SVM learning methods are easy-to-

\footnote{This research was funded in part by Israel Science Foundation grant 1308/12 and partly by United States NCI grant CA142538. The first author is thankful to the Friday Machine Learning Lab at the Biostatistics Department, University of North Carolina at Chapel Hill, for many helpful discussions and suggestions.}

\textit{Keywords and phrases:} Empirical risk minimization, Support vector machines, Variable selection, Recursive feature elimination.
compute techniques that enable estimation under weak or no assumptions on the distribution (see Steinwart and Christmann, 2008). SVM learning methods, which we review in detail in Section 2, are a collection of algorithms that attempt to minimize a regularized version of the empirical risk over some reproducing kernel Hilbert space (RKHS) with respect to some loss function. The standard SVM decision function typically utilizes all the input variables. Hence, when the input dimension is large, it can suffer from the so-called ‘Curse of Dimensionality’ (Hastie et al., 2001). A procedure for variable selection is thus of importance to obtain a more intelligible solution with improved efficiency. The advantages of variable selection are multi fold: it increases the generalized performance of the learning, it clarifies the causal relationship in the input-output space, and results in reduced cost of data collection and storage and better computational properties.

One of the earliest works on variable selection in SVM was formulated by Guyon et al. (2002). Guyon et al. developed a backward elimination procedure based on recursive computation of the SVM learning function, known widely as recursive feature elimination (RFE). The RFE algorithm performs a recursive ranking of a given set of features. At each recursive step of the algorithm, it calculates the change in the RKHS norm of the estimated SVM function after deletion of each of the features remaining in the model, and removes the one with the lowest change in such norm. The process thus performs an implicit ranking of the features and can even be generalized to remove chunks of features at each step of recursion. A number of modified approaches have been developed since then, inspired by RFE (see Rakotomamonjy, 2003; Aksu et al., 2010; Aksu, 2012). Although there is no dearth of rich literature on RFE for SVMs, the theoretical properties of it have never been studied. The arguments for RFE have mostly been heuristic and its ability to produce successful data-driven performances in simulated or real-life settings. A key reason behind this lack of theory is the absence of a well-established framework for building, justifying, and collating the theoretical foundation of such a feature elimination method. This paper aims at building such a framework and validating RFE as a theoretically sound procedure for feature elimination in SVMs.

Developing a theoretical structure for RFE is challenging. At each stage of the feature elimination process, we move down to a ‘lower dimensional’ feature space and the functional spaces need to be adjusted to cater to the appropriate version of the problem in these subspaces. Euclidean spaces, for example, as well as many specialized functional classes admit a nested structure in this regard, but as we will see later, this is not true in general. As mentioned before, SVM attempts to minimize
the empirical regularized risk within an RKHS of functions. Starting with a given RKHS, one daunting task is re-defining the functional space so that it retains the premises of the original space (i.e. admits the reproducing structure) and that these spaces remain cognate to one another. The basis for the theory on RFE depends heavily on correctly specifying these pseudo-subspaces, and a contribution of this paper is to formulate a way to do this.

Another contribution of this paper is a modification of the criterion for deletion and ranking of features in Guyon et al.’s RFE to enable theoretical consistency. Here we develop a ranking of the features based on the lowest difference observed in the regularized empirical risk after removing each feature from the existing model. The definition of RFE used here can thus be generalized to the much broader yet simpler setting of empirical risk minimization where we can apply the same idea to the empirical risk. This can thus serve as a useful starting point for more in-depth theoretical analysis of feature elimination in SVM. While Guyon et al.’s RFE tends to rely on the penalization criterion in the SVM objective function for ranking features, our approach is risk-based, in that we utilize the entire objective function for ranking. The heuristic reasoning behind this is that if any of the features do not contribute to the model at all, the increase in the regularized risk will be inconsequential.

In this paper, we show that the modified RFE is asymptotically consistent in finding the ‘correct’ feature space both for SVMs and empirical risk minimization (ERM) under reasonable regularity conditions. Although these regularity conditions are true for most of the relevant problems at hand, we show through appropriate examples that consistency results for RFE might fail in general, and for correct utilization of RFE as a consistent tool for feature elimination in SVMs, we need these regularity conditions to hold. The notion of consistency in such a context has not been defined previously. This paper also aims at positing a basis for which such results are meaningful. A comprehensive statistical analysis of SVMs can be found in Steinwart and Chirstmann (2008) (hereafter abbreviated SC08) which is used in this paper to develop the concept of consistency for RFE in the context of feature elimination in SVM and ERM. We give an in-depth analysis of a few case studies, including the setting of risk minimization in linear models and SVM for classification with a Gaussian RBF kernel, to show how the results developed here can be applied to specific examples. We also provide some simulation results to validate our theoretical conclusions and discuss how to utilize the proposed deletion criteria to select the important features in a given
setting.

While RFE is a popular and simple method for variable selection, several other methods do exist in the context of feature elimination in SVMs. RFE is a classic example of a \textit{wrapper} that uses the learning method itself to score feature subsets. Alternative wrapper-based selection methods have also been formulated for feature elimination in SVMs (Weston et al., 2001; Chapelle et al., 2002). Other basic types of variable selection techniques include \textit{filters} that select subsets of the feature space as a pre-processing step or \textit{embedded methods} that construct the learning algorithm in a way to include feature elimination as an in-built phenomenon. Filters have been used for feature elimination in SVMs in many previous works (see for example Mladenic et al., 2004; Peng et al., 2005). Embedded variable selection methods include redefining the SVM training to include sparsity (Weston et al., 2003; Chan et al., 2007). For example, Bradley and Mangasarian (1998) suggested the use of the $l_1$ penalty to encourage feature sparsity. Zhu et al. (2003) suggested an algorithm to compute the solution path for this $l_1$-norm SVM efficiently. Other methods include introducing different penalty functions like the SCAD penalty (Zhang et al., 2006), the $l_q$ penalty (Liu et al., 2007), a combination of $l_0$ and $l_1$ penalty (Liu and Wu, 2007), the elastic net (Wang et al., 2006), the $f_\infty$ norm (Zou and Yuan, 2006), and using a penalty functional in the framework of the smoothing spline ANOVA (Zhang, 2006). Although these alternative methods appear to perform well in practice, RFE still remains the most widely used methodology for feature selection in support vector machines due to its simplicity and generality.

In Section 2, we give a short preliminary background for empirical risk minimization and support vector machines. In Section 3 we present the proposed version of RFE for ERM and SVM. In Section 4 we discuss the concept of feature elimination in these frameworks. In Section 5 we give the necessary assumptions for RFE in empirical risk minimization and support vector machines and provide a short discussion on the meaningfulness of these assumptions in varied situations. The associated consistency results for RFE are given in Section 6. In Section 7 we discuss our results in some known settings of ERM and SVM. In Section 8 we provide some simulation results to demonstrate how RFE works and how it can be used in intelligent selection of features. A discussion is provided in Section 9, detailed proofs are given in the Appendix, and the resources for the necessary codes are given in Supplement A.
2. Preliminaries. We start off with some preliminaries and define the notations that we will follow for the rest of the paper. We also give a brief introduction to support vector machines and empirical risk minimization.

2.1. Empirical risk minimization. Empirical risk minimization (ERM) is a general setting of many supervised learning problems.

Let the input space \((\mathcal{X}, \mathcal{A})\) be measurable, such that \(\mathcal{X} \subseteq B \subseteq \mathbb{R}^d\) where \(B\) is an open Euclidean ball centered at 0, and \(\mathcal{Y} \subseteq \mathbb{R}\) and let \(P\) be a measure on \(\mathcal{X} \times \mathcal{Y}\). A function \(L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \mapsto [0, \infty]\) is called a loss function if it is measurable. We say that a loss function is convex if \(L(x, y, \cdot)\) is convex for every \(x \in \mathcal{X}\) and \(y \in \mathcal{Y}\). A loss function is called locally Lipschitz continuous with Lipschitz local constant \(c_L(\cdot)\) if for every \(a > 0\),

\[
\sup_{x \in \mathcal{X}, y \in \mathcal{Y}} |L(x, y, s) - L(x, y, \hat{s})| < c_L(a) |s - \hat{s}|, \quad s, \hat{s} \in [-a, a].
\]

The loss function \(L\) is said to be Lipschitz continuous if there is a constant \(c_L\) such that \(c_L(a) \leq c_L \forall a \in \mathbb{R}\).

For any measurable function \(f : \mathcal{X} \mapsto \mathbb{R}\) we define the L-risk of \(f\) with respect to the measure \(P\) as \(\mathcal{R}_{L,P}(f) = E_P[L(X, Y, f(X))]\). The Bayes Risk \(\mathcal{R}_{L,P}^*\) with respect to the loss function \(L\) and the measure \(P\) is defined as \(\inf_f \mathcal{R}_{L,P}(f)\), where the infimum is taken over the set of all measurable functions, \(\mathcal{L}_0(\mathcal{X}) = \{f : \mathcal{X} \mapsto \mathbb{R}, \ f \text{ is measurable}\}\). A function \(f_P^*\) that achieves this infimum is called a Bayes decision function.

Let \(\mathcal{F} \subseteq \mathcal{L}_0(\mathcal{X})\) be a non-empty functional space, and \(L\) be any loss function. Let

\[
(1) \quad f_{P,F} = \arg \min_{f \in \mathcal{F}} E_P[L(X, Y, f(X))] = \arg \min_{f \in \mathcal{F}} \mathcal{R}_{L,P}(f)
\]

be the minimizer of infinite-sample risk within the space \(\mathcal{F}\). Define the minimal risk within the space \(\mathcal{F}\) as \(\mathcal{R}_{L,P,F}^* = \mathcal{R}_{L,P}(f_{P,F})\). Define the empirical risk \(\mathcal{R}_{L,D}\) as \(\mathcal{R}_{L,D}(f) = \mathbb{P}_n(L(X, Y, f(X))) = \frac{1}{n} \sum_{i=1}^{n} L(X_i, Y_i, f(X_i))\).

A learning method whose decision function \(f_{D,F}\) satisfies \(\mathcal{R}_{L,D}(f_{D,F}) = \inf_{f \in \mathcal{F}} \mathcal{R}_{L,D}(f)\) for all \(n \geq 1\) and \(D = \{(X_1, Y_1), \ldots, (X_n, Y_n)\} \in (\mathcal{X} \times \mathcal{Y})^n\) is called empirical risk minimization (ERM) with respect to \(L\) and \(\mathcal{F}\).
2.2. **Support vector machines.** The results developed for SVM in this paper are valid not only for classification, but also for regression under some general assumptions on the output space \(Y\), however throughout this paper we would refer to all these versions as SVM.

Let \(H\) be an \(\mathbb{R}\)-Hilbert function space over \(X\), then a function \(k: X \times X \rightarrow \mathbb{R}\) is called a reproducing kernel of \(H\) if \(k(\cdot, x) \in H\) for all \(x \in X\) and has the reproducing property that \(f(x) = \langle f, k(\cdot, x) \rangle\) for all \(f \in H\) and all \(x \in X\). The space is called a Real-valued Reproducing Kernel Hilbert Space (RKHS) over \(X\) if for all \(x \in X\), the Dirac functional \(\delta_x: H \rightarrow \mathbb{R}\) defined by \(\delta_x(f) := f(x)\) is continuous for all \(f \in H\) (For more details refer to chapter 4 of SC08).

Let \(L\) be a convex, locally Lipschitz continuous loss function and \(H\) be a separable RKHS of a measurable kernel \(k\) on \(X\). Let \(D = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}\) be a set of \(n\) i.i.d observations drawn according to the probability measure \(P\) and fix a \(\lambda > 0\). Define the empirical SVM decision function as

\[
(2) \quad f_{D,\lambda,H} = \arg\min_{f \in H} \lambda \|f\|_H^2 + \mathcal{R}_{L,D}(f),
\]

where \(\mathcal{R}_{L,D}(f)\) is the empirical risk defined as before.

For a given \(\lambda\), define the SVM learning method \(\mathcal{L}\) as the map \((X \times Y)^n \times X \rightarrow \mathbb{R}\) defined by \((D, x) \mapsto f_{D,\lambda,H}(x)\), for all \(n \geq 1\). We say that a learning method \(\mathcal{L}\) is measurable if it is measurable for all \(n\) with respect to the minimal completion of the product \(\sigma\)-field on \((X \times Y)^n \times X\). Lemma 6.23 of SC08, under the conditions given in Section 2.2 above yields that the corresponding SVM that produces the decision functions \(f_{D,\lambda,H}\) for \(D \in (X \times Y)^n\) is a measurable learning method for all \(\lambda > 0\) and for all \(n \geq 1\). The maps \(D \mapsto f_{D,\lambda,H}\) mapping \((X \times Y)^n\) to \(H\) are measurable. Since Lemma 2.11 of SC08 shows that the map \(H \times X \mapsto \mathbb{R}\) defined by \((f, x) \mapsto f(x)\) is measurable, we therefore obtain measurability of \((D, x) \mapsto f_{D,\lambda,H}(x)\).

Define \(f_{P,\lambda,H} = \arg\min_{f \in H} \lambda \|f\|_H^2 + \mathcal{R}_{L,P}(f)\) and define the approximation error

\[
(3) \quad A_H^2(\lambda) = \lambda \|f_{P,\lambda,H}\|_H^2 + \mathcal{R}_{L,P}(f_{P,\lambda,H}) - \inf_{f \in H} \mathcal{R}_{L,P}(f).
\]

2.3. **Entropy Numbers.** For \((T, d)\) a metric space and for any integer \(n \geq 1\), the \(n\)-th entropy number of \((T, d)\) is defined as

\[
(4) \quad e_n(T, d) := \inf \left\{ \epsilon > 0 : \exists s_1, \ldots, s_{2^n-1} \in T \text{ such that } T \subset \bigcup_{i=1}^{2^n-1} B_d(s_i, \epsilon) \right\}
\]
A NOTION OF CONSISTENCY FOR RFE IN SVMS AND ERMS

where \( B_d(s, \epsilon) \) is the ball of radius \( \epsilon \) centered at \( s \), with respect to the metric \( d \). Let \( S : E \mapsto F \) be a bounded linear operator between normed spaces \( E \) and \( F \), then we write \( e_n(S) = e_n(SB_E, \| \cdot \|_F) \), where \( B_E \) is the unit ball in \( E \).

Note: If we have \( \|k\|_{\infty} < \infty \) for a given kernel \( k \), Lemma 4.23 of SC08 implies that every \( f \in H(k) \) is bounded which further implies that \( H \subseteq L_{\infty}(X) \), where \( L_{\infty}(X) := \{ f : X \mapsto \mathbb{R}, \|f\|_{\infty} < \infty \} \).

3. Feature Elimination Algorithm. The original RFE (Recursive Feature Elimination) Method was proposed for SVMs by Guyon et al. (2002). The feature elimination procedure version we propose here is similar to the one in Guyon et al. except for the elimination criterion. While Guyon et al. use the criterion Hilbert space norm \( \lambda \|f\|_H^2 \) to eliminate features recursively, we use the entire objective function including the regularized Hilbert Space norm along with the empirical risk. Hence while Guyon’s RFE is only applicable in analyses involving SVM, the modified RFE that we propose here can be used in ERM as well.

3.1. The Algorithms. The RFE was originally developed for support vector machines, hence we provide the algorithm for SVM first. The definition for ERM follows similarly. First we define some related concepts. A detailed discussion on these will be given in Section 4.

Definition 1. For any set of indices \( J \subseteq \{1, 2, ..., d\} \) and a given functional space \( F \), define \( F^J = \{ g : g = f \circ \pi^J, \forall f \in F \} \), where \( \pi^J \) is the projection map taking \( x \) to \( x^J \) \( (x, x^J \in \mathbb{R}^d) \), such that \( x^J \) is produced from \( x \) by replacing elements of \( x \) indexed in the set \( J \), by zero.

We define the space \( \mathcal{X}^J = \{ \pi^J(x) : x \in \mathcal{X} \} \), such that \( \pi^J : \mathcal{X} \mapsto \mathcal{X}^J \) is a surjection.

Definition 2. For a given RKHS \( H \) indexed by a kernel \( k \) and for a given \( J \), define \( H^J = H_{k_{o \pi^J}}(\mathcal{X}) \), where \( k_{o \pi^J}(x, y) := k(\pi^J(x), \pi^J(y)) \).

Now we are ready to provide the algorithm. Assume the support vector machine framework, where we are given an RKHS \( H \) with respect to a kernel \( k \).

Algorithm 3. Start off with \( J \equiv [] \) empty and let \( Z = [1, 2, ..., d] \).
STEP 1: In the $k^{th}$ cycle of the algorithm choose dimension $i_k$ for which

\[
    i_k = \arg \min_{i \in \mathbb{Z} \setminus J} \lambda \left\| f_{D,\lambda, H^\cup\{i\}} \right\|_{H^\cup\{i\}}^2 + R_{L,D} \left( f_{D,\lambda, H^\cup\{i\}} \right) - \lambda \left\| f_{D,\lambda, H^\cup\{i\}} \right\|_{H^\cup\{i\}}^2 - R_{L,D} \left( f_{D,\lambda, H^\cup\{i\}} \right).
\]

STEP 2: Update $J = J \cup \{i_k\}$. Go to STEP 1.

Continue this until the difference

\[
    \min_{i \in \mathbb{Z} \setminus J} \lambda \left\| f_{D,\lambda, H^\cup\{i\}} \right\|_{H^\cup\{i\}}^2 + R_{L,D} \left( f_{D,\lambda, H^\cup\{i\}} \right) - \lambda \left\| f_{D,\lambda, H^\cup\{i\}} \right\|_{H^\cup\{i\}}^2 - R_{L,D} \left( f_{D,\lambda, H^\cup\{i\}} \right)
\]

becomes larger than a pre-determined quantity $\delta_n$.

Now for an empirical risk minimization framework with respect to a given functional space $F$, Algorithm 3 can be modified to match the setting of ERM.

**Algorithm 4.** Replace the regularized empirical risk $\lambda \left\| f_{D,\lambda, H^\cup\{i\}} \right\|_{H^\cup\{i\}}^2 + R_{L,D} \left( f_{D,\lambda, H^\cup\{i\}} \right)$ in Algorithm 3, (defined for a given set of indices $J$) by the empirical risk $R_{L,D} \left( f_{D,F^J} \right)$.

3.2. Cycle of RFE. We define ‘cycle’ of the RFE algorithm as the number of ‘dimensions’ deleted in one step of the algorithm. The algorithms in 3.1 has cycle = 1. But one can define it for cycles of value greater than 1 in which case one deletes chunks of dimensions at a time, equal to the size of the cycle. It can also be defined adaptively such that in different runs of the algorithm the cycle sizes are different. The theoretical results derived in this paper will hold for cycles of any size. Hence, for the sake of simplicity, we set the cycle size to 1.

4. Functional Spaces on Lower Dimensional Domains. The aim of this section is to provide a detailed reasoning behind Definitions 1–2 in Section 3.1.

4.1. Feature Elimination in ERM. Suppose we have a functional class $F \subseteq \mathcal{L}_\infty(\mathcal{X})^1$, where $\mathcal{X}$ is as defined in Section 2 and let our goal be to find a function $f$ within the functional class $F$ which minimizes an empirical criterion (like empirical risk in ERM). But if the dimension $d$ of the input space is too large, it might lead to more complex solutions when in fact a simpler solution might be good enough. Now suppose that the minimizer of the appropriate infinite-sample version of the empirical criterion (like risk or expected loss in case of ERM and SVM) with respect to

\[\text{Note that the loss functions we consider in this paper (unless otherwise mentioned) are convex and locally Lipschitz with } R_{L,P}(0) < \infty, \text{ and hence by (2.11) and Proposition 5.27 of SC08, we have } R^*_{L,P,\mathcal{L}_\infty(\mathcal{X})} = R^*_{L,P}. \text{ Hence instead of } \mathcal{L}_0(\mathcal{X}) \text{ it suffices to consider the smaller subspace } \mathcal{L}_\infty(\mathcal{X}).\]
the probability measure $P$ on $\mathcal{X} \times \mathcal{Y}$ and the functional class $\mathcal{L}_\infty(\mathcal{X})$, actually resides in $\mathcal{L}_\infty(\mathcal{X}^*)$ where $\mathcal{X}^*$ is a lower dimensional version of the given input space $\mathcal{X}$. Then to avoid over fitting it is necessary that we try to find the empirical minimizer in a suitably defined lower dimensional version of $\mathcal{F}$. We define the lower dimensional adaptations of the original functional space as in Definition 1.

First note that $\mathcal{X}^J$ may not be a subspace of $\mathcal{X}$, because for any $x \in \mathcal{X}$, $\pi^J(x)$ may not be contained in $\mathcal{X}$. Note that the assertion holds trivially for any Euclidean open ball $B$ centered at 0, and from Section 2 we have that $\mathcal{X} \subseteq B$, for some $B \subset \mathbb{R}^d$. Hence we will assume that the functional space $\mathcal{F}$ can be sufficiently redefined as $\mathcal{F}_B$, where the domain of functions in $\mathcal{F}_B$ is $B$ instead of $\mathcal{X}$, such that $\mathcal{F}_B|_\mathcal{X} = \mathcal{F}$. This makes the functional classes $\mathcal{F}^J$ well-defined, and unless otherwise mentioned, we will assume from hereon that $\mathcal{X}^J \subseteq \mathcal{X}$ for all possible $J$.

Note also that $\mathcal{F}^J$ may not be a subspace of $\mathcal{F}$. Although it is more desirable for these functional classes to accept a nested structure between each other, so that as we go down from a space to its lower dimensional version (that is, from $\mathcal{F}^{J_1}$ to $\mathcal{F}^{J_2}$ where $J_1 \subseteq J_2$, we can have the simple relation that $\mathcal{F}^{J_2} \subseteq \mathcal{F}^{J_1}$), it does not hold in general.

We now provide a few results that connect the space $\mathcal{F}$ with its lower dimensional versions. Note that our definition trivially implies that $\mathcal{F}^J|_{\mathcal{X}^J} = \{f|_{\mathcal{X}^J} : f \in \mathcal{F}\} = \mathcal{F}|_{\mathcal{X}^J}$. Now if we define $\mathcal{L}_\infty^J(\mathcal{X}) = \{f \circ \pi^J : f \in \mathcal{L}_\infty(\mathcal{X})\}$, then Lemma 5 says that $\mathcal{L}_\infty^J(\mathcal{X}^J) \equiv \mathcal{L}_\infty^J(\mathcal{X})|_{\mathcal{X}^J}$ is isomorphic to the space $\mathcal{L}_\infty(\mathcal{X}^J)$. Lemma 6 below shows that by defining the functional classes in this way, many of the nice properties of the functional class $\mathcal{F}$ are carried forward to the $\mathcal{F}^J$s. The proofs can be found in Appendix A.1 and A.2.

**Lemma 5.** $\mathcal{L}_\infty^J(\mathcal{X}^J) = \mathcal{L}_\infty(\mathcal{X}^J)$.

**Lemma 6.** Let $\mathcal{F} \subseteq \mathcal{L}_\infty(\mathcal{X})$ be a non-empty functional subspace. Then for any $J \subseteq \{1, 2, \ldots, d\}$,

1. If $\mathcal{F}$ is dense in $\mathcal{L}_\infty(\mathcal{X})$, then $\mathcal{F}^J$ is dense in $\mathcal{L}_\infty^J(\mathcal{X})$.
2. If $\mathcal{F}$ is compact, then so is $\mathcal{F}^J$.
3. $e_i(\mathcal{F}^J, \|\cdot\|_\infty) \leq e_i(\mathcal{F}, \|\cdot\|_\infty)$, $\forall i \geq 1$ where $e_i(\mathcal{F}, \|\cdot\|_\infty)$ is the $i^{th}$ entropy number of the set $\mathcal{F}$ with respect to the $\|\cdot\|_\infty$-norm as defined in Section 2.3.
4.2. Feature Elimination in SVM. In empirical risk minimization problems our primary focus is the empirical risk, whereas in the case of support vector machines we concentrate on the regularized version of the empirical risk, \( \lambda \| f \|_H^2 + R_{L,D}(f) \). The minimization is typically computed over an RKHS \( H \), that is, our objective is to find \( f_{D,\lambda,H} \equiv \arg\min_{f \in H} \lambda \| f \|_H^2 + R_{L,D}(f) \). The regularization term \( \lambda \| f \|_H^2 \) is used to penalize functions \( f \) with a large RKHS norm. Complex functions \( f \in H \) which model too closely the output values in the training data set \( D \), tend to have large \( H \)-norms (Refer to Exercise 6.7 in SC08 for a clear motivation). Again we assume that \( X \subseteq B \subset \mathbb{R}^d \) where \( B \) is an open Euclidean ball centered at 0. We will also assume that we can sufficiently re-define the RKHS \( H \) as \( H_B \), such that the domain of the functions in \( H_B \) is the Euclidean open ball \( B \) instead of \( X \). So we can extend the domain of the kernel \( k \) of the RKHS \( H \) from \( X \times X \) to \( B \times B \) and from here onwards we assume \( X \subseteq B \subset \mathbb{R}^d \). The usual way that we defined the lower dimensional functional spaces in the previous section may not be sufficient here mainly because in SVM, the minimization is computed over an RKHS, and the properties of RKHSs dictate a lot of the statistical analyses. Hence we need to find a way to define them so that these spaces are RKHSs as well.

First we review some properties of RKHS:

1. The ‘\( H_{\text{pre}} \)’ space for an RKHS \( H \) with kernel \( k \) is defined as \( H_{\text{pre}} := \left\{ \sum_{i=1}^n \alpha_i k(\cdot, x_i) : n \in \mathbb{N}, \alpha_1, \ldots, \alpha_n \in \mathbb{R}, x_1, \ldots, x_n \in X \right\} \). \( H \) is the completion of the space \( H_{\text{pre}} \) (See Theorem 4.16 of SC08 for details).

2. Let \( S \) be any set and \( \varphi : S \mapsto X \) be a map. Let \( k : X \times X \mapsto \mathbb{R} \) be the kernel on \( X \). Then define the map \( k \circ \varphi : S \times S \mapsto \mathbb{R} \) as, \( k \circ \varphi(s,t) = k(\varphi(s), \varphi(t)) \). Observe that \( k \circ \varphi \) is a kernel on \( S \) (Paulsen, 2009, Proposition 5.13).

The next theorem gives a natural relationship between the RKHS \( H(k) \) on \( X \) and the RKHS \( H(k \circ \varphi) \) on \( S \). Also when \( S \) is a subset of \( X \) and \( \varphi \) is the inclusion \( \text{id} \) map of \( S \) into \( X \), then the kernel \( k \circ \varphi \) is the restriction of the kernel \( k \) on \( S \times S \).

**Theorem 7.** Let \( X \) and \( S \) be two sets and let \( k : X \times X \mapsto \mathbb{R} \) be a kernel function on \( X \) and let \( \varphi : S \mapsto X \) be a function. Then \( H(k \circ \varphi) = \{ f \circ \varphi : f \in H(k) \} \), and for \( g \in H(k \circ \varphi) \) we have that \( \| g \|_{H(k \circ \varphi)} = \inf \{ \| f \|_H(k) : g = f \circ \varphi \} \).

See Paulsen (2009) for a proof of Theorem 7.
Now let $\mathcal{X}_0$ be a subset of $\mathcal{X}$ and $k^{(0)}(x, y)$ be the restriction of the kernel $k$ on $\mathcal{X}_0$ and $H_{k^{(0)}}(\mathcal{X})$ be the RKHS admitting $k^{(0)}(x, y)$ as its reproducing kernel and $H_k(\mathcal{X})$ be the RKHS with its reproducing kernel $k(x, y)$. Then by the above theorem and defining $\varphi$ to be the inclusion id map from $\mathcal{X}_0$ to $\mathcal{X}$, we have $H_{k^{(0)}}(\mathcal{X}_0) = \{f|_{\mathcal{X}_0} : f \in H_k(\mathcal{X})\}$ and $\|g\|_{H_{k^{(0)}}} = \min\{\|f\|_{H_k} : f|_{\mathcal{X}_0} = g\}$ for $g \in H_{k^{(0)}}(\mathcal{X}_0)$.

For a given RKHS $H_k(\mathcal{X})$, we can now define these new functional spaces in the following ways:

- **Def 1** Projection of the Functional Space: We can define it as we did in the previous section. So define $H_J^J$ on $\mathcal{X}$ as $H_J^J = H_J(\mathcal{X}) = \{f \circ \pi^J_c : f \in H_k(\mathcal{X})\}$.

- **Def 2** Projection of the kernel: $H_J^J$ defined on $\mathcal{X}$ as $H_J^J = H_{k \circ \pi^J_c}(\mathcal{X})$. Note that by defining them like this, the new spaces that we obtain are all RKHSs on $\mathcal{X}$.

From the discussion below Theorem 7 and in Def 2 we have,

$$H_J^J|_{\mathcal{X}^J} = H_{k \circ \pi^J_c}(\mathcal{X}^J) = \{f|_{\mathcal{X}^J} : f \in H_k(\mathcal{X})\}.$$  

(6)

Also note from Def 1 that

$$H_J^J|_{\mathcal{X}^J} = H^J(\mathcal{X}^J) = \{f \circ \pi^J_c|_{\mathcal{X}^J} : f \in H_k(\mathcal{X})\} = \{f|_{\mathcal{X}^J} : f \in H_k(\mathcal{X})\}.$$  

(7)

So we see that restrictions of both of these functional spaces to $\mathcal{X}^J$ are the same and the restriction space is itself an RKHS on $\mathcal{X}^J$. Also note trivially that $H_{k \circ \pi^J_c}(\mathcal{X}^J) \equiv H_{k \circ \pi^J_c}(\mathcal{X})$ and hence from now onwards we would refer to the space $H_{k \circ \pi^J_c}(\mathcal{X})$ as simply $H^J$.

Next we redefine Lemma 6 for the RKHSs. The proofs are similar and hence omitted:

**Lemma 8.** Let $H \subset L_\infty(\mathcal{X})$ be a non-empty RKHS on $\mathcal{X}$. Then for any $J \subset \{1, 2, \ldots, d\}$,

1. If $H$ is dense in $L_\infty(\mathcal{X})$, then $H^J$ is dense in $L_\infty(\mathcal{X}^J)$.
2. If the $\| \cdot \|_\infty$ closure $\overline{B_H}$ of the unit ball $B_H$ is compact, then so is $\overline{B_{H^J}}$.
3. If $H$ is separable, then so is $H^J$.
4. $e_i(id : H^J \mapsto L_\infty(\mathcal{X}^J)) \leq e_i(id : H \mapsto L_\infty(\mathcal{X}))$, where $e_i(id : H \mapsto L_\infty(\mathcal{X}))$ is the $i$-th entropy number of the unit ball $B_H$ of the RKHS $H$, with respect to the $\| \cdot \|_\infty$-norm.

In order to provide a heuristic understanding of the importance of the above projection spaces in feature selection, we give an alternative definition of lower dimensional versions of the input
space. First, define the map $\sigma^J : \mathbb{R}^d \rightarrow \mathbb{R}^{|J|}$ such that for $x = \{x_1, \ldots, x_d\} \in \mathbb{R}^d$, $\sigma^J(x) = \{x_{j_{\text{min}}}, \ldots, x_{j_{\text{max}}}\} \in \mathbb{R}^{|J|}$. So $\sigma^J(x)$ is the $|J|$ dimensional vector containing only those elements of $x$ which are given in the index set $J$. Hence we can define the deleted input space $\mathcal{X}^{-J}$ as, $\mathcal{X}^{-J} := \{\sigma^J(x) = \{x_{j_{\text{min}}}, \ldots, x_{j_{\text{max}}}\} : x \in \mathcal{X}\}$.

Consider the set up of Theorem 7, with $\mathcal{X} \equiv \mathcal{X}^J$, and $\mathcal{S} \equiv \mathcal{X}^{-J}$. Consider the restricted kernel $k^J$ on $\mathcal{X}^J$ with $k^J(x, y) = k(x, y)$ for all $x, y \in \mathcal{X}^J$. Now for any $y \in \mathcal{X}^{-J}$ define the map $\varphi \equiv \phi^J_d : \mathcal{X}^{-J} \mapsto \mathcal{X}^J$ as $\phi^J_d(y) = \pi^J(x)$ for any $x \in \mathcal{X}$ satisfying $y = \sigma^J(x)$. Or in other words the map $\phi^J_d$ takes an element from the deleted space, fills in the gaps with zeros and returns an element from the projected space. Note then that $\phi^J_d$ is a bijection, and hence the spaces $\mathcal{X}^J$ and $\mathcal{X}^{-J}$ are isomorphic to each other.

Hence from Theorem 7, we see that $k^J \circ \phi^J_d$ is a kernel defined on $\mathcal{X}^{-J}$ and with the corresponding RKHS $H_{k^J \circ \phi^J_d}$ on $\mathcal{X}^{-J}$. Suppose that instead of $\mathcal{X}$, our input space is $\mathcal{X}^{-J}$. We want to know when can we define a kernel $k^{-J}$ on $\mathcal{X}^{-J}$ such that it is the natural abridgment of the kernel $k$ on $\mathcal{X}$ (in the sense of being able to define it on deleted vectors) and we want to know if there exists a natural connection between $H_{k^{-J}}(\mathcal{X}^{-J})$ and $H_{k^J \circ \phi^J_d}(\mathcal{X}^{-J})$ in those cases.

The motivation for the definition of $k^{-J}$ stems from previous works on feature elimination in Support Vector Machines. The Recursive Feature Elimination procedure developed in Guyon et al. (2002) and subsequently revisited and modified in Rakotomamonjy (2003) starts off with a given input space $\mathcal{X}$ and eliminates features using a weight criteria recursively computed by re-training the SVM on the lower dimensional spaces $\mathcal{X}^{-J}$. From their discussion, it is seen that if the Gram matrix of the training vectors $\{x_1, \ldots, x_n\}$ is given by $\{k(x_k, x_j)\}_{k,j=1}^n$, then the Gram matrix of the training vectors $\{x_i^{-1}, \ldots, x_i^{-i}\}$ after deleting a particular variable say $\mathcal{X}_i$ is taken to be $\{k^{-i}(x_k, x_j)\}_{k,j=1}^n$ where $k^{-i}(x_k, x_j) = k(x_k^{-i}, x_j^{-i})$. This clearly takes into account the assumption that the kernel $k$ can be defined on deleted vectors as well, that is, $k$ is well defined for any pair of vectors $x$ and $y$ where $x, y \in \mathbb{R}^{d_0}$ and $d_0 \leq d$. This is clearly not true in general for any kernel $k$ on $\mathbb{R}^d$. So we prefer to work with the projected spaces $\mathcal{X}^J$ instead of the deleted spaces $\mathcal{X}^{-J}$ as this is more general. But we will show in the discussions below that in most practical cases as discussed in Guyon et al. (2002), and Rakotomamonjy (2003), many of the kernels that we work with satisfy an intrinsic relationship between $k^{-J}$ and $k^J \circ \phi^J_d$. Hence in those cases it is appropriate to work with either of the two setups.
4.2.1. **Kernels in Statistical Learning.** Most popular kernels in statistical learning can be categorized into three main groups: translation invariant kernels, kernels originating from generative models (like Jaakkola and Haussler, Watsons) and dot product kernels (see Smola et al., 2000). In this paper we restrict our attention to only translation invariant and dot product kernels.

- **Translation Invariant Kernels:** A translation invariant Kernel satisfies $k(x, y) = g(x - y)$. The class of translation invariant kernels also includes Radial Kernels which satisfy $k(x, y) = g(\|x - y\|^2)$, where $\|x\|$ is the usual Euclidean norm of vector $x$ in its correct dimension.

- **Dot-Product Kernels:** A dot-product kernel satisfies $k(x, y) = g(\langle x, y \rangle)$, where $\langle x, y \rangle$ is the standard inner product between vectors $x$ and $y$ in their correct dimension.

**Lemma 9.** For Radial Kernels and Dot Product Kernels, $k^{-J} = k^J \circ \phi_d^J$.

The proof is simple and therefore omitted.

Also note that for kernels defined on weighted norms, $(k(x, y) = g(\|x - y\|_W))$ where $\|x - y\|_W := (x - y)'W(x - y)$, with $W$ being a positive $d \times d$ diagonal matrix), the above condition is also satisfied.

4.2.2. **Universal Kernels.** A continuous kernel $k$ on a compact metric space $\mathcal{X}$ is called universal if the RKHS $H_k(\mathcal{X})$ is dense in $C(\mathcal{X})$, i.e., for every function $g \in C(\mathcal{X})$ and all $\epsilon > 0$ there exists an $f \in H_k(\mathcal{X})$ such that $\|f - g\|_\infty \leq \epsilon$ (where $C(\mathcal{X})$ denotes the set of continuous functions from $\mathcal{X} \mapsto \mathbb{R}$). From Proposition 5.29 of SC08 we see that if $\mathcal{X}$ is a compact metric space with $H_k(\mathcal{X})$, the RKHS of a universal kernel $k$ on $\mathcal{X}$, $P$ a distribution on $\mathcal{X} \times \mathcal{Y}$ and $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \mapsto [0, \infty)$ a convex, locally Lipschitz continuous loss with $\mathcal{R}_{L,P}(0) < \infty$, then we have that $\mathcal{R}_{L,P,H_k} = \mathcal{R}_{L,P}^*$.

Universal kernels produce particularly large RKHSs.

4.2.3. **Universality of Kernels.** For this we refer our Readers to Micchelli et al. (2006) where the notion of universality for most of the special types of kernels are discussed in details (including dot product and radial kernels). However we state two results on radial kernels here to show that under quite weak assumptions, all of the non trivial radial kernels are universal.

(RK1) **Representation of Radial Kernels:** Schoenberg (1938) showed that a function $k(x, y) : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}$
$\mathbb{R}^d \rightarrow \mathbb{R}$ defined as

$$k(x, y) := g(||x - y||^2), \quad x, y \in \mathbb{R}^d,$$

where $\| \cdot \|$ is the usual Euclidean norm, is a valid kernel on $\mathbb{R}^d \times \mathbb{R}^d$ for all $d \in \mathbb{N}$ iff there exists a finite Borel measure $\mu$ on $\mathbb{R}_+$ such that for all $t \in \mathbb{R}_+$,

$$g(t) := \int_{\mathbb{R}_+} e^{-t\sigma} d\mu(\sigma).$$

All kernels of this type are not universal. Indeed, the choice of a measure concentrated only at $\sigma = 0$ gives a kernel $k$ that is identically constant and therefore it is not universal. The next result shows that this is the only exceptional case.

(RK2) Universal of Radial Kernels: If the measure $\mu$ in (8) is not concentrated at zero then the radial kernel $k$ in (9) is universal (For proof see Micchelli et al. (2006)).

4.3. Notion of risk in Lower Dimensional Versions of the Input Space. Note that the functional space $\mathcal{F}^J$ (and equivalently RKHS $H^J$) is defined on the entire input space $\mathcal{X}$ and not only on $\mathcal{X}^J$. So we can define risk for a function $f_J \in \mathcal{F}^J$ (or $f_J \in H^J$) for the entire input space $\mathcal{X}$ and not just for $\mathcal{X}^J$. Hence for a probability distribution $P$ on $\mathcal{X} \times \mathcal{Y}$, define $R_{L,P}(f_J)$ as $R_{L,P}(f_J) = \int_{\mathcal{Y}} \int_{\mathcal{X}} L(y, x, f_J(x)) P(x, y) dx dy$. This means that we can compare the risk of functions in different lower dimensional versions of the original functional space.

5. Assumptions for RFE and Their Implications. In this section we discuss the assumptions needed for consistency of RFE for both ERM and SVM. We then discuss validity of these assumptions under practical settings.

5.1. Assumptions. Consider the setting of risk minimization (regularized or non regularized) with respect to a given functional space $\mathcal{F}$ (which are typically RKHSs in case of SVM). Our main aim is to provide a framework where the modified recursive feature elimination method we proposed earlier is consistent in finding the correct lower dimensional subspace of the input space, and the assumptions required for this are:
(A1). Let $J$ be a subset of $\{1, \ldots, d\}$. Let the function $f_{P,\mathcal{F}}$ minimize risk within the space $\mathcal{F}^J$ with respect to the probability distribution $P$ on $\mathcal{X} \times \mathcal{Y}$. Here, we define $\mathcal{F}^\emptyset = \mathcal{F}$. We assume that there exists a non-trivial $J_*$ that is $J_* \neq \emptyset$ and $|J_*| = d_0$ that satisfies the criterion that for any pair $(d_1, d_2)$ such that $d_1 \leq d_2 \leq d_0$, $\exists J_{d_1}$ and $J_{d_2}$ such that $J_{d_1} \subseteq J_{d_2} \subseteq J_*$ with $|J_{d_1}| = d_1$ and $|J_{d_2}| = d_2$ such that $\mathcal{R}_{L,P,\mathcal{F}^{J_*}} = \mathcal{R}_{L,P,\mathcal{F}^{J_{d_1}}} = \mathcal{R}_{L,P,\mathcal{F}^{J_{d_2}}}$.

In other words, Assumption (A1) says that there exists a ‘path’ from the original input space $\mathcal{X}$ to the correct lower dimensional space $\mathcal{X}^{J_*}$ in the sense of equality of the minimized risk within the functional spaces $\mathcal{F}^J$s along this ‘path’. So there exists a sequence of indices $\mathcal{J}$ from $J_{\text{start}} = \emptyset$ to $J_{\text{end}} = J_*$, where $\mathcal{J} := \{J_{\text{start}} \equiv J_1, J_2, \ldots, J_\ast \equiv J_{\text{end}} \} : J_1 \subseteq J_2 \subseteq \cdots \subseteq J_{\text{end}}, |J_i| = |J_{i-1}| + 1\}$, such that $\mathcal{R}_{L,P,\mathcal{F}}^{J_*}$ is the same for all $J \in \mathcal{J}$. Note that $\mathcal{J}$ may not be unique and there might be more than one path leading to $\mathcal{X}^{J_*}$. Also note that $J_*$ may not be unique in general, but any one of them would work for our purpose. So we will assume it to be unique in this paper.

(A2). Let $\mathcal{J}_1, \mathcal{J}_2, \ldots, \mathcal{J}_N$ be the exhaustive list of such paths from $\mathcal{X}$ to $\mathcal{X}^{J_*}$, and let $\mathcal{J} := \bigcup_{i=1}^N \mathcal{J}_i$.

There exists $\epsilon_0 > 0$ such that whenever $J \notin \mathcal{J}$, $\mathcal{R}_{L,P,\mathcal{F}}^{J_*} \geq \mathcal{R}_{L,P,\mathcal{F}}^{\mathcal{J}_i} + \epsilon_0$.

In Section 6 we will show that Assumptions (A1) and (A2) are sufficient for a recursive feature elimination algorithm like RFE to work (in terms of consistency). Here we try to show the necessity of Assumption (A1) in order for a well-defined recursive feature elimination algorithm to work.

5.2. Necessity of existence of a path in (A1).

**Example 10.** Consider the empirical risk minimization framework. Let $X = [-1,1]^2$ and let $Y = 0$. Let $X_1 \sim \mathcal{U}$ where $\mathcal{U}$ is some distribution on $[-1,1]$ and $X_2 \equiv -X_1$. Let the functional space $\mathcal{F}$ be $\{c(X_1 + X_2), c > 0\}$. Let the loss function be the squared error loss, i.e., $L(x, y, f(x)) = (y - f(x))^2$. By Definition 1, $\mathcal{F}^{(1)} = \{cX_2, c > 0\}$ and $\mathcal{F}^{(2)} = \{cX_1, c > 0\}$ and $\mathcal{F}^{(1,2)} = \{0\}$. We see that $\mathcal{R}_{L,P}(f_{P,\mathcal{F}}) = \mathcal{R}_{L,P}(f_{P,\mathcal{F}^{(1,2)}}) = 0$ but both $\mathcal{R}_{L,P}(f_{P,\mathcal{F}^{(1)}})$ and $\mathcal{R}_{L,P}(f_{P,\mathcal{F}^{(2)}}) \neq 0$. Hence even if the correct low-dimensional functional space may have minimized risk the same as that of the original functional space, if there does not exist a path going down to that space, the recursive algorithm will not work. Note that the minimizer of the risk belongs to $\mathcal{F}^{(1,2)}$ but there is no path from $\mathcal{F}$ to $\mathcal{F}^{(1,2)}$, in the sense of (A1).
5.3. Necessity of Equality in (A1). It would appear that for the algorithm to work, we don’t have to necessarily work with equalities along the path and that we can relax (A1) to include inequalities as well. Suppose we redefine (A1) as (A1*), where the equality of minimized risk along the path is replaced by ‘≤’. So now we assume that minimized risk is not necessarily constant along the path, but that it does not increase. We show below that under this modified assumption, our recursive search algorithm might fail to find the correct lower dimensional subspace of the input space.

**Example 11.** Let \( Y \sim U(-1, 1) \) and \( X \subset \mathbb{R}^3 \) such that \( Y = X_3 = X_2 + 1 = X_1 - 1 \). Let \( \mathcal{F} = \{ c_1 X_1 + c_2 X_2 + c_3 X_3, c_1, c_2, c_3 \geq 1 \} \), and let the loss function be squared error loss. Now by definition, \( \mathcal{F}^{(1)} = \{ c_2 X_2 + c_3 X_3, c_2, c_3 \geq 1 \} \), \( \mathcal{F}^{(2)} = \{ c_1 X_1 + c_3 X_3, c_1, c_3 \geq 1 \} \), \( \mathcal{F}^{(3)} = \{ c_2 X_2 + c_1 X_1, c_1, c_2 \geq 1 \} \), \( \mathcal{F}^{(1, 2)} = \{ c_3 X_3, c_3 \geq 1 \} \), \( \mathcal{F}^{(1, 3)} = \{ c_2 X_2, c_2 \geq 1 \} \), \( \mathcal{F}^{(2, 3)} = \{ c_1 X_1, c_1 \geq 1 \} \), and \( \mathcal{F}^{(1, 2, 3)} = \{ 0 \} \).

By simple calculations, we see that \( R_{L,P,F}^* = R_{L,P,F}^{(1)} = R_{L,P,F}^{(2)} = 4/3 \), \( R_{L,P,F}^{(3)} = R_{L,P,F}^{(1, 2, 3)} = 1/3 \), \( R_{L,P,F}^{(1, 3)} = R_{L,P,F}^{(2, 3)} = 1 \) and \( R_{L,P,F}^{(1, 2)} = 0 \). Note that the correct dimensional subspace of the input space is \( X^{(1, 2)} \) and there exists paths leading to this space via \( X \rightarrow X^{(1)} \rightarrow X^{(1, 2)} \) since \( R_{L,P,F}^* = R_{L,P,F}^{(1)} > R_{L,P,F}^{(1, 2)} \) or via \( X \rightarrow X^{(2)} \rightarrow X^{(1, 2)} \) since \( R_{L,P,F}^* = R_{L,P,F}^{(2)} > R_{L,P,F}^{(1, 2)} \) in the sense of Assumption (A1*). But there also exists the blind path \( X \rightarrow X^{(3)} \) since \( R_{L,P,F}^* > R_{L,P,F}^{(3)} \) which does not lead to the correct subspace. Hence the recursive search in this case may not be guaranteed to lead to the correct subspace.

Hence equality in (A1) guarantees that the recursive search will never select an important dimension \( j \in J_\ast \) for redundancy because then the Assumption (A2) would be violated. Hence the equality in (A1) will ensure that we will follow a path recursively to the correct input space \( X^{J_\ast} \).

5.4. Validity of the Assumptions in Practical Situations. In this section, we discuss the validity of the assumptions in 5.1, with respect to practical situations of risk minimization.

In ERM, our main aim is to find a function \( f \) within a class \( \mathcal{F} \) that minimizes empirical risk within that class. Choice of the functional space \( \mathcal{F} \) is important as it determines a fine balance between complexity of the solution (see discussion in 4) on one hand and finding a function that has risk close to the Bayes risk on the other. Often the spaces we consider for minimization satisfy properties that make the assumptions in Section 5.1 fairly natural. For SVM, the choice of the
RKHS $H$ is just as important as it was in choosing a functional class $F$ in ERMs. Again, in most practical situations, the RKHS $H$ will satisfy some properties that would make Assumptions (A1) and (A2) quite standard conditions for feature elimination.

5.4.1. Nested spaces in empirical risk minimization. Often the space $F$ we consider for ERM is such that for any $J$, $F^J \subseteq F$, that is, it admits the nested property. So for any $J_1, J_2 \in \{1, 2, \ldots, d\}$ with $J_1 \subseteq J_2$ we have $F^{J_2} \subseteq F^{J_1}$. This means we also have nested inequalities in the form of $R_{L,P,F}^{J_1} \leq R_{L,P,F}^{J_2}$ for such $J_1$ and $J_2$. One example is the linear combination space where the coefficients are allowed to take values in a compact interval containing 0, $F = \{f(x_1, \ldots, x_d) = \sum_i a_i x_i : |a_i| \leq M, M < \infty\}$. In these cases, simple observation shows that Assumption (A1) translates to saying that there exists a minimizer $f_{P,F}$ which minimizes infinite-sample risk in $F$, satisfying the criterion that $f_{P,F} \in F^{J^{\ast}}$ which further implies that $f_{P,F} \in F_J$ for any $J \subseteq J^{\ast}$. Then the results of Section 5.1 imply that for any $J \subseteq J^{\ast}$, $R_{L,P,F}^{J^{\ast}} = R_{L,P,F}^{J}$ and for any $J_0 \not\subseteq J^{\ast}$, $R_{L,P,F}^{J_0} \geq R_{L,P,F}^{J}$ + $\epsilon_0$ for $\epsilon_0$ as defined in (A2).

Even if the space $F$ does not satisfy the nested criterion, we can create the nested structure for feature elimination by considering the unions of these spaces. Noting that $F \equiv F^0$, we can create them as follows:

$$\tilde{F}^J = \bigcup_{J \subseteq J^{\ast} \subseteq \{1, \ldots, d\}} F^{J^{\ast}}.$$  (10)

It can be seen that the properties of $F$ and $F^J$s with respect to Lemma 6 are carried forward in our new definitions too. If $F$ is dense in $L_\infty(C)$, by Lemma 6 we have $F^J$ dense in $L_\infty^J(C)$ for any $J$, which implies that $\tilde{F}^J$ is dense in $L_\infty^J(C)$. If $F$ is compact, Lemma 6 implies that $F^J$ is compact for any $J$, hence $\tilde{F}^0$ and $\tilde{F}^J$s are compact too, since the unions only include finitely many terms. If we have $e_i(F, \|\cdot\|_\infty) < \infty$ then again by Lemma 6, $e_i(\tilde{F}, \|\cdot\|_\infty) < \infty$ and $e_i(\tilde{F}^J, \|\cdot\|_\infty) < \infty$ for all $J$.

5.4.2. Nested RKHSs. Unfortunately, in general, RKHSs need not be nested in each other. As we discussed in ERM, given any RKHS $H$, we cannot create unions of RKHSs to use them in learning, because unions of RKHSs may not be RKHSs. Here we discuss cases where the naturally occurring RKHSs are in fact nested within each other. We will see that dot-product kernels actually have this property. To see this, let us consider a dot-product kernel $k$ such that $k(x, y) = g(\langle x, y \rangle)$
where $\langle \cdot, \cdot \rangle$ is the usual Euclidean inner-product. Now let us consider the pre-RKHSs $H_{\text{pre}}$ and $H^J_{\text{pre}}$. We show here that $H^J_{\text{pre}} \subseteq H_{\text{pre}}$ which will imply that $H^J \subseteq H$. For this, take $f \in H^J_{\text{pre}}$ which implies that $f$ can be written as $f(\cdot) = \sum_{i=1}^{n} \alpha_i k^J(\cdot, x_i)$ for $n \in \mathbb{N}$, $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$, $x_1, \ldots, x_n \in \mathcal{X}$. Hence,

$$f(\cdot) = \sum_{i=1}^{n} \alpha_i k_j(\cdot, x_i) = \sum_{i=1}^{n} \alpha_i k \left( \pi^{J^\infty}(\cdot), \pi^{J^\infty}(x_i) \right) = \sum_{i=1}^{n} \alpha_i g \left( \langle \pi^{J^\infty}(\cdot), \pi^{J^\infty}(x_i) \rangle \right)
= \sum_{i=1}^{n} \alpha_i g \left( \langle \cdot, \pi^{J^\infty}(x_i) \rangle \right) = \sum_{i=1}^{n} \alpha_i k \left( \cdot, \pi^{J^\infty}(x_i) \right)$$

Noting that $\pi^{J^\infty}(x_1), \ldots, \pi^{J^\infty}(x_n) \in \mathcal{X}$, we have that $f \in H_{\text{pre}}$. In a similar way, we can show that for any $J_1 \subseteq J_2$, $H^{J_2} \subseteq H^{J_1}$.

In the case of RKHSs produced by dot-product kernels (and potentially other RKHSs satisfying the nested property), the implications of (A1) and (A2) will be the same as discussed before (see the above section on nested spaces in 5.4.1), and we omit the details here.

5.4.3. Dense spaces in empirical risk minimization. Another wide class of function spaces we typically consider in ERM are dense spaces. So if we have that $\mathcal{F}$ is dense in $\mathcal{L}_\infty(\mathcal{X})$, Lemma 6 gives us that $\mathcal{F}^J$ is dense in $\mathcal{L}_\infty^J(\mathcal{X})$ for any $J \in \{1, 2, \ldots, d\}$. First note that for $J_1$ and $J_2$ with $J_1 \subseteq J_2$, we trivially have $\mathcal{L}_\infty^{J_2}(\mathcal{X}) \subseteq \mathcal{L}_\infty^{J_1}(\mathcal{X}) \subseteq \mathcal{L}_\infty(\mathcal{X})$. This also means that for such $J_1$ and $J_2$ we have $\mathcal{R}_{L,P,F,J_2}^* \geq \mathcal{R}_{L,P,F,J_1}^*$. Now ‘denseness’ does not necessarily imply ‘nestedness’, but we have the ‘almost nested’ property in the sense that for any $g \in \mathcal{F}^{J_2}$, and for any $\epsilon > 0$, $\exists f_\epsilon \in \mathcal{F}^{J_1}$ with $\|f_\epsilon - g\|_\infty \leq \epsilon$. This means that for $J_* \in \mathcal{X}$ as defined in (A1), $\exists \{f_n\} \in \mathcal{F}$, such that $f_n(x) \rightarrow f_{L,P,F,J_*}(x)$. Hence in terms of (A1) since the loss functions we consider are locally Lipschitz continuous, by Lemma 2.17 of SC08 we have $\mathcal{R}_{L,P}(f_n) \rightarrow \mathcal{R}_{L,P,F,J_*}^* = \mathcal{R}_{L,P,F}^*$.

Also note that (A1) implies that for any $J \subseteq J_*$, $\mathcal{R}_{L,P,F,J}^* \geq \mathcal{R}_{L,P,F,J_*}^* = \mathcal{R}_{L,P,F}^*$. Also note for any $J_0 \not\subseteq J_*$, $\mathcal{R}_{L,P,F,J}^* \geq \mathcal{R}_{L,P,F,J_0}^* + \epsilon_0$ for $\epsilon_0$ as defined in (A2).

5.4.4. Dense RKHSs. Most of the times the RKHS we would consider for SVMs will also be dense in $\mathcal{L}_\infty(\mathcal{X})$. Note that the properties of these dense spaces with respect to our Assumptions (A1) and (A2) as discussed above will remain the same here as well. All universal kernels produce RKHSs that are dense in $\mathcal{L}_\infty(\mathcal{X})$ with respect to convex, locally Lipschitz continuous losses and
RK1 and RK2 imply that all non-trivial radial kernels share this property as well. Hence in most situations, our Assumptions (A1) and (A2) are a natural way to define a premise that necessitates feature elimination.

6. Consistency Results for RFE. In this section we show that Algorithms 3 and 4 defined in Section 3.1 are consistent in finding the correct feature space under the assumptions in 5.1. We will refer often to SC08 for the theory and results developed in their text.

6.1. Theoretical Results. Before we state the main results of this section, we note down some necessary conditions for general ERMs, and for SVMs. We start with ERM:

(B1). Let $L$ be a convex locally Lipschitz continuous loss function.
(B2). Let $\mathcal{F} \subset \mathcal{L}_\infty(\mathcal{X})$ be non-empty and compact.
(B3). There exists $M > 0$ satisfying $\|f\|_\infty \leq M$, $f \in \mathcal{F}$.
(B4). There exists $B > 0$ satisfying $L(x, y, f(x)) \leq B$, $(x, y) \in \mathcal{X} \times \mathcal{Y}$, $f \in \mathcal{F}$.
(B5). Assume that for fixed $n \geq 1$, there exists constants $a \geq 1$ and $p \in (0, 1)$ such that
$$E_{D_X \sim P^n_X}\epsilon_i(\mathcal{F}, L_\infty(D_X)) \leq ai^{-\frac{1}{p}}, \quad i \geq 1;$$
where $E_{D_X \sim P^n_X}$ is defined as the expectation with respect to the product measure $P^n_X$ under the assumption that the input data $D_X \equiv \{X_1, \ldots, X_n\}$ are i.i.d. copies of $X \sim P_X$.

We now define conditions for SVMs:

(C1). Let $P$ be a probability measure on $\mathcal{X} \times \mathcal{Y}$ where the input space $\mathcal{X}$ is a valid metric space.
(C2). Let $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \mapsto [0, \infty]$ be a convex locally Lipschitz continuous loss function satisfying $L(x, y, 0) \leq 1$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$.
(C3). Let $H$ be the separable RKHS of a measurable kernel $k$ on $\mathcal{X}$ with $\|k\|_\infty \leq 1$.
(C4). Assume that for fixed $n \geq 1$, there exists constants $a \geq 1$ and $p \in (0, 1)$ such that
$$E_{D_X \sim P^n_X}\epsilon_i(id : H \mapsto L_\infty(D_X)) \leq ai^{-\frac{1}{p}}, \quad i \geq 1.$$
(C5). For a sample size of $n$ we choose a $\lambda_n \in [0, 1]$ such that $\lambda_n \to 0$ and $\lim_{n \to \infty} \lambda_n n = \infty$.
(C6). $\exists c > 0$ and $\beta \in (0, 1)$ such that $A_2J(\lambda) \leq c\lambda^\beta$ for any $J$ and for all $\lambda \geq 0$ (where $A_2J(\lambda) \equiv A_2^HJ(\lambda)$).

Conditions (B1) – (B5) will be used for proving the results for ERM, while (C2) – (C6) will be used for proving the results for SVM. Note that these are standard assumptions that are typically
used in the statistical analyses of empirical risk minimization and support vector machines (Refer to Chapter 6 and 7 of SC08).

Also note that the conditions $L(x, y, 0) \leq 1$ in (C2), and $\|k\|_\infty \leq 1$ for the kernel $k$ in (C3) are assumed for simplicity and might be too restrictive in some settings, but equivalent conditions like $L(x, y, 0) \leq M$ and $\|k\|_\infty \leq k_{\text{sup}}$ for constants $M, k_{\text{sup}} > 1$ are good enough for the proofs and will result in bounds differing from the ones derived here only up to some constants.

**Theorem 12.** Assume (B1) – (B5). For $\delta_n^{-1} = O(n^{\frac{1}{2}})$ with $\delta_n \to 0$, we have the following:

1. The Recursive Feature Elimination Algorithm for empirical risk minimization, defined for $\{\delta_n\}$ given above, will find the correct lower dimensional subspace of the input space with probability tending to 1.

2. The function chosen by the algorithm achieves the best risk within the original functional space $\mathcal{F}$ asymptotically.

**Theorem 13.** Assume (C1) – (C6). If we take $\delta_n^{-1} = O(n^{\frac{\beta}{2\beta+1}})$ with $\delta_n \to 0$, then we have the following:

1. The Recursive Feature Elimination Algorithm for support vector machines, defined for $\{\delta_n\}$ given above, will find the correct lower dimensional subspace of the input space with probability tending to 1.

2. The function chosen by the algorithm achieves the best risk within the original RKHS $H$ asymptotically.

We will give a detailed proof of Theorem 12 later and it will be seen that proof of Theorem 13 will be similar. But first, we provide a few relevant results which we will need for proving these main results. We start off with the following lemma.

**Lemma 14.** Let $(\mathcal{F}, \| \cdot \|_\mathcal{F})$ be a separable functional space, such that the metric $\| \cdot \|_\mathcal{F}$ dominates pointwise convergence. Also we assume sup $\|f\|_\mathcal{F} \leq C$ for some $C < \infty$ for all $f \in \mathcal{F}$. Let $L$ be a convex, locally Lipschitz loss function such that $L(x, y, f(x)) \leq B$ for some $B < \infty$ for all $f \in \mathcal{F}$. Also assume that for fixed $n \geq 1$, $\exists$ constants $a \geq 1$ and $p \in (0, 1)$ such that $\mathbb{E}_{D_X \sim p_X^n} e_i(\mathcal{F}, L_\infty(D_X)) \leq a i^{-\frac{p}{2}}$, $i \geq 1$. Then, we have with probability greater than or equal to
1 - e^{-\tau},
\[
\sup_{f \in \mathcal{F}} |R_{L,P}(f) - R_{L,D}(f)| \leq 2B \sqrt{\frac{2\tau}{n}} + \frac{10B\tau}{3n} + 4 \max \left\{ C_1(p)c_L(C)p + B^{1-p}n^{-\frac{1}{2}}, C_2(p)c_L(C)^{\frac{2p}{1+p}}B^{\frac{1}{1+p}}n^{-\frac{1}{1+p}} \right\}.
\]

See Appendix A.3 for a proof. We now provide the necessary results for ERM.

**Proposition 15.** Assume conditions (B1) - (B5). For all measurable ERMs and all \( \epsilon > 0, \tau > 0, \) and \( n \geq 1, \) and for \( J_1, J_2 \in \mathcal{J} \) such that \( J_1 \subseteq J_2 \subseteq J_* \), we have

\[
P^n \left( D \in (\mathcal{X} \times \mathcal{Y})^n : |R_{L,D}(f_{D,J_2}) - R_{L,D}(f_{D,J_1})| < 12B \sqrt{\frac{2\tau}{n}} + \frac{20B\tau}{n} + 24K_1 \left( \frac{a^{2p}}{n} \right)^{\frac{1}{2}} \right) > 1 - 2e^{-\tau},
\]

where \( K_1 := \max \left\{ \frac{B}{4}, C_1(p)c_L(C)p + B^{1-p}n^{-\frac{1}{2}}, C_2(p)c_L(C)^{\frac{2p}{1+p}}B^{\frac{1}{1+p}} \right\} \).

The proof can be found in Appendix A.4.

Note that compactness of \( \mathcal{F} \) implies compactness of \( \mathcal{F}^J \) for any \( J \) by Lemma 6. Compactness is important because along with continuity of \( R_{L,D} : \mathcal{L}_\infty(\mathcal{X}) \to [0, \infty) \), it ensures the existence of an empirical risk minimizer. Also compactness of \( \mathcal{F}^J \) implies that it is a closed and separable subset of \( \mathcal{L}_\infty(\mathcal{X}) \) for each \( J \). Hence Lemma 6.17 of SC08 shows that there exists a measurable ERM for both classes \( \mathcal{F}^{J_1} \) and \( \mathcal{F}^{J_2} \). Now note that for any \( J \), the quantities \( R_{L,D}(f_{D,J}), R_{L,P}^*(f_{D,J}) \) and \( R_{L,P}(f_{D,J}) \) are all less than or equal to \( B \). Note also that,

\[
|R_{L,D}(f_{D,J}) - R_{L,P}^*(f_{D,J})| \leq |R_{L,D}(f_{D,J}) - R_{L,P}(f_{D,J})| + |R_{L,P}(f_{D,J}) - R_{L,P}^*(f_{D,J})|
\]

\[
\leq \sup_{f \in \mathcal{F}^J} |R_{L,P}(f) - R_{L,D}(f)| + 2 \sup_{f \in \mathcal{F}} |R_{L,P}(f) - R_{L,D}(f)|.
\]

Consequently we obtain the following two corollaries:

**Corollary 16.** Assume the conditions of Proposition 15. For any \( J \) and all measurable ERMs and all \( \epsilon > 0, \tau > 0, \) and \( n \geq 1, \) we have that

\[
P^n \left( D \in (\mathcal{X} \times \mathcal{Y})^n : |R_{L,D}(f_{D,J}) - R_{L,P}^*(f_{D,J})| < 6B \sqrt{\frac{2\tau}{n}} + \frac{10B\tau}{n} + 12K_1 \left( \frac{a^{2p}}{\lambda n} \right)^{\frac{1}{2}} \right) > 1 - e^{-\tau},
\]

where \( K_1 \) is as before. Additionally if \( J \in \mathcal{J} \), we can replace \( R_{L,P}^*(f_{D,J}) \) in the above inequality by \( R_{L,P}^*(f_{D,J}) \).
**Corollary 17.** Oracle Inequality for ERM: Assume the conditions of Proposition 15. For any \( J \) and all \( \epsilon > 0, \tau > 0, \) and \( n \geq 1, \) we have with \( P^n \) probability \( 1 - e^{-\tau}, \)

\[
\mathcal{R}_{L,P} \left( \hat{f}_{D,\mathcal{J}} \right) - \mathcal{R}^*_L < 4B \sqrt{\frac{2\tau}{n}} + \frac{20B\tau}{3n} + 8K_1B^{1-p} \left( \frac{a^{2p}}{\lambda^{p\epsilon n}} \right)^{\frac{1}{2}},
\]

where \( K_1 \) is as before.

We now present similar results for SVMs. As before we start off the main proposition:

**Proposition 18.** Assume conditions (C1) – (C4). For fixed \( \lambda > 0, \epsilon > 0, \tau > 0, \) and \( n \geq 1, \) and for \( J_1, J_2 \in \mathcal{J} \) such that \( J_1 \subseteq J_2 \subseteq J_*, \) we have with probability \( P^n \) not less than \( 1 - 2e^{-\tau}, \)

\[
\left( f_{D,\lambda,H_2} \right) - \mathcal{R}^*_L < A_2^{J_1}(\lambda) + B_2^{J_2}(\lambda) + 12B \sqrt{\frac{2\tau}{n}} + \frac{20B\tau}{3n} + 24K_1B^{1-p} \left( \frac{a^{2p}}{\lambda^{p\epsilon n}} \right)^{\frac{1}{2}},
\]

where \( A_2^{J_1}(\lambda) \) and \( A_2^{J_2}(\lambda) \) are the approximation errors for the two separate RKHS classes \( H^{J_1} \) and \( H^{J_2}, \) \( B := cL(\lambda^{-1/2})\lambda^{-1/2} + 1, \) and \( K_2 := \max \left\{ B^p/4, C_1(p)cL(\lambda^{-1/2})^{2p}, C_2(p)cL(\lambda^{-1/2}) \right\} \) is a constant depending only on \( B, p \) and the Lipschitz constant \( cL(\lambda^{-1/2}). \)

See Appendix A.5 for a detailed proof of Proposition 18.

Note that since \( B \geq 1 \) and \( K \geq B^p/4, \) we have that if \( a^{2p} > \lambda^{p\epsilon n}, \)

\[
\left( f_{D,\lambda,H_j} \right) - \mathcal{R}^*_L \leq \mathcal{R}_L(0) + \mathcal{R}_{L,P}(0) \leq 2
\]

\[
< 3B \leq 12KB^{1-p} \left( \frac{a^{2p}}{\lambda^{p\epsilon n}} \right)^{\frac{1}{2}}.
\]

Similarly, since \( B \geq 1 \) and \( K \geq B^p/4, \) we have for \( a^{2p} > \lambda^{p\epsilon n}, \)

\[
\left( f_{D,\lambda,H_j} \right) - \mathcal{R}^*_L \leq \mathcal{R}_L(0) + \mathcal{R}_{L,P}(0) \leq 1 + 2B
\]

\[
\leq 8KB^{1-p} \left( \frac{a^{2p}}{\lambda^{p\epsilon n}} \right)^{\frac{1}{2}}.
\]

Now note that for any \( J, \) we have

\[
\left( f_{D,\lambda,H_j} \right) - \mathcal{R}^*_L \leq \mathcal{R}_L(0) + \mathcal{R}_{L,P}(0) \leq 1 + 2B
\]

\[
< 8KB^{1-p} \left( \frac{a^{2p}}{\lambda^{p\epsilon n}} \right)^{\frac{1}{2}}.
\]
Consequently, we obtain two corollaries for SVMs, similar to Corollaries 16 and 17:

**Corollary 19.** Assume the conditions of Proposition 18. For any $J$ and all $\epsilon > 0$, $\tau > 0$, and $n \geq 1$, we have with $P^n$ probability $> 1 - e^{-\tau}$,

$$\left| \lambda \left\| f_{D,\lambda,H,J} \right\|_2^2 + R_{L,D} \left( f_{D,\lambda,H,J} \right) - R_{L,P,H,J}^* \right| < A^2_2(\lambda) + 6B \sqrt{\frac{2\tau}{n}} + 10B \frac{\tau}{n} + 12K_2 B^{1-p} \left( \frac{a^2}{\lambda^p n} \right)^{\frac{1}{2}},$$

where $K_2$ is as before. Additionally, if $J \in \tilde{J}$, we can replace $R_{L,P,F,J}^*$ in the above inequality by $R_{L,P,F,J}^*$.

**Corollary 20.** Oracle Inequality for SVM: Assume the conditions of Proposition 18. For any $J$ and all $\epsilon > 0$, $\tau > 0$, and $n \geq 1$, we have with $P^n$ probability $> 1 - e^{-\tau}$,

$$\lambda \left\| f_{D,\lambda,H,J} \right\|_2^2 + R_{L,P} \left( f_{D,\lambda,H,J} \right) - R_{L,P,H,J}^* < A^2_2(\lambda) + 4B \sqrt{\frac{2\tau}{n}} + \frac{20B \tau}{3n} + 8K_2 B^{1-p} \left( \frac{a^2}{\lambda^p n} \right)^{\frac{1}{2}},$$

where $K_2$ is as before.

Proposition 15 and Corollaries 16, 17 are intended for ERM and will be used in proving Lemma 21 which in turn will aid in proving Theorem 12. Similarly, Proposition 18 and Corollaries 19, 20 developed for SVM will be used to prove the following Lemma 22, that will set up the premise for proving Theorem 13.

We now provide Lemma 21 for ERM:

**Lemma 21.** Assume the conditions of Proposition 15. Then the following statements hold:

i. For $J_1, J_2 \in \tilde{J}$ and $J_1 \subset J_2 \subset J_*$, $\exists \{(\epsilon_n) > 0\} \rightarrow 0$ such that we have with probability greater than $1 - 2e^{-\tau}$, $R_{L,D} \left( f_{D,F,J_2} \right) \leq R_{L,D} \left( f_{D,F,J_1} \right) + \epsilon_n$.

ii. For $J_1 \in \tilde{J}$, $J_2 \notin \tilde{J}$ such that $J_1 \subset J_2$, $\exists \{\epsilon_n\} > 0$ and $\epsilon_n \rightarrow \epsilon_0 > 0$, such that we have with probability greater than $1 - 2e^{-\tau}$, $R_{L,D} \left( f_{D,F,J_2} \right) > R_{L,D} \left( f_{D,F,J_1} \right) + \epsilon_n$.

iii. Oracle Property for RFE in ERM: For a given $J \subseteq \{1, \ldots, d\}$ the infinite-sample risk of the function $f_{D,F,J}$, $R_{L,P} \left( f_{D,F,J} \right)$, converges in measure to $R_{L,P,F,J}^*$ (and hence to $R_{L,P,F,J}^*$ if $F$ is dense in $L_\infty (\lambda)$) iff $J \in \tilde{J}$.
(C5). To explicitly establish rates for our algorithm we further assume that the bound on the approximation error $A_2^J(\lambda)$ is as given in (C6).

Lemma 22. Assume the conditions of Proposition 18. Also assume (C5) and (C6). Then the following statements hold:

i. For $J_1, J_2 \in \tilde{J}$ such that $J_1 \subseteq J_2 \subseteq J_*$, $\exists (\{\epsilon_n \}) > 0 \rightarrow 0$ such that we have with probability greater than $1 - 2e^{-\tau}$,

$$\lambda_n \|f_D,\nu, H_{j_2} \|^2_{H_{j_2}} + R_{L,D}(f_D,\nu, H_{j_2}) \leq \lambda_n \|f_D,\nu, H_{j_1} \|^2_{H_{j_1}} + R_{L,D}(f_D,\nu, H_{j_1}) + \epsilon_n.$$ 

ii. For $J_1 \in \tilde{J}$ and $J_2 \notin \tilde{J}$ and for $J_1 \subset J_2$, $\exists \{\tilde{\epsilon}_n \} > 0$ and $\tilde{\epsilon}_n \rightarrow \epsilon_0 > 0$, such that we have with probability greater than $1 - 2e^{-\tau}$,

$$\lambda_n \|f_D,\nu, H_{j_2} \|^2_{H_{j_2}} + R_{L,D}(f_D,\nu, H_{j_2}) \geq \lambda_n \|f_D,\nu, H_{j_1} \|^2_{H_{j_1}} + R_{L,D}(f_D,\nu, H_{j_1}) + \tilde{\epsilon}_n.$$ 

iii. Oracle Property for RFE in SVM: The infinite-sampled regularized risk for the empirical solution $f_D,\nu, H_{j_*}$, $\lambda_n \|f_D,\nu, H_{j_*} \|^2_{H_{j_*}} + R_{L,P}(f_D,\nu, H_{j_*})$ converges in measure to $\mathcal{R}_{L,P,H}$ (and hence to $\mathcal{R}_{L,P}^*$ if the RKHS $H$ is dense in $L_\infty(X)$) iff $J \in \tilde{J}$.

The proof of Lemma 22 in given in Appendix A.6. Lemma 21 follows similarly and hence we only discuss briefly the results that we obtain from Lemma 21 in Appendix A.7.

We are now ready to prove Theorems 12 and 13. Since the proofs are similar we only provide the proof for Theorem 12 and discuss the changes required for the proof of Theorem 13.

6.2. Proof of Theorem 12.

Proof. (1) Let $\mathcal{X}^{j_*}$ be the correct input space and $J_*$ be the correct set of dimensions to be removed with $|J_*| = d_0$. To prove the first part of Theorem 12, we show that, starting with the input space $\mathcal{X}$, the probability that we reach the space $\mathcal{X}^{j_*}$ is 1 asymptotically. First let us assume that there exists only one correct ‘path’ from $\mathcal{X}$ to $\mathcal{X}^{j_*}$. Let $J^\circ$ be that correct path and $J^\circ = \{j_0^\circ \equiv \{\}, j_1^\circ, \ldots, j_{d_0}^\circ \equiv J_*\}$. For the sake of notational ease, we use the convention that

$$\mathcal{R}_{L,D}(f_D,\mathcal{X}) = \mathcal{R}_{L,D,F}^*.$$ 

So from (32) in Appendix A.7 we have, $\mathcal{R}_{L,D,F}^* - \mathcal{R}_{L,D,F}^{j_{i+1}} < \epsilon_n$ with probability at least $1 - 2e^{-\tau}$, where $\epsilon_n = 12B\sqrt{2\tau n^{-1/2}} + 20B\tau n^{-1} + 24K_1a^{-1}p n^{-1/2}$. Now let $J_{i+1} \neq J_{i+1}$ be any other $J$.
such that \( J_i^o \subset J_{i+1} \) with \( \| J_{i+1} \| = \| J_i^o \| + 1 \), then from the proof of Lemma 21 in Appendix A.7 we have \( R^*_{L,D,F,J_{i+1}} - R^*_{L,D,F,J_i^o} > \epsilon_0 - \epsilon_n \), with probability at least \( 1 - 2e^{-\tau} \). Now if we choose \( \tau = o(n) \) with \( \tau \to \infty \), then \( \delta_n \equiv \epsilon_n \) satisfies the conditions that \( \delta_n^{-1} = O(n^{1/2}) \) with \( \delta_n \to 0 \). Now since \( \epsilon_0 \) is a fixed constant, \( \exists N_{\epsilon_0} > 0 \) such that \( \forall n \geq N_{\epsilon_0}, 2\delta_n < \epsilon_0 \). Hence, without loss of generality, we assume throughout the remainder of the proof that \( n \geq N_{\epsilon_0} \). Then we have the condition that

\[
R^*_{L,D,F,J_{i+1}} - R^*_{L,D,F,J_i^o} > \delta_n \text{ with probability at least } 1 - 2e^{-\tau}.
\]

Then,

\[
P(\text{`RFE finds the correct dimensions'}) \geq P(\text{`RFE follows the path } J^o \text{ to the correct dimension space'})
= P(J_0 := J_0^o, J_1 := J_0^o, \ldots, J_{d_0} := J_{d_0}^o, J_{d_0+1} := \emptyset)
= P(J_0 := J_0^o) P(J_1 := J_1^o \mid J_0^o) \cdots P(J_{d_0} := J_{d_0}^o \mid J_0^o, \ldots, J_{d_0-1}^o) P(J_{d_0+1} := \emptyset \mid J_0^o, \ldots, J_{d_0}^o),
\]

where `\( J_{d_0+1} := \emptyset \)` means the algorithm stops at that step. Note that \( P(J_0 := J_0^o) = 1 \) and then observe,

\[
P(J_{i+1} := J_{i+1}^o \mid J_0^o, \ldots, J_i^o)
= P(J_{i+1} := J_{i+1}^o \mid J_0^o) \quad (\because \{ J_0^o, \ldots, J_{i-1}^o \} \text{ have already been removed from the model})
\]

\[
= P \left( R^*_{L,D,F,J_{i+1}^o} - R^*_{L,D,F,J_i^o} \leq \delta_n, R^*_{L,D,F,J_{i+1}^o} - R^*_{L,D,F,J_i^o} < R^*_{L,D,F,J_{i+1}^o} - R^*_{L,D,F,J_i^o} \forall J_{i+1}^* \neq J_{i+1}^o \right)
= P \left( R^*_{L,D,F,J_{i+1}^o} - R^*_{L,D,F,J_i^o} \leq \delta_n, \delta_n < R^*_{L,D,F,J_{i+1}^o} - R^*_{L,D,F,J_i^o} \forall J_{i+1}^* \neq J_{i+1}^o \right)
\geq P \left( R^*_{L,D,F,J_{i+1}^o} - R^*_{L,D,F,J_i^o} \leq \delta_n \right) - \sum_{J_{i+1}^* \neq J_{i+1}^o} P \left( R^*_{L,D,F,J_{i+1}^o} - R^*_{L,D,F,J_i^o} > \delta_n \right)
\geq 1 - 2e^{-\tau} - 2(d-i-1)e^{-\tau} = 1 - 2(d-i)e^{-\tau}.
\]

Also see that

\[
P(J_{d_0+1} := \emptyset \mid J_0^o, \ldots, J_{d_0}^o) = P \left( R^*_{L,D,F,J_{d_0+1}^o} - R^*_{L,D,F,J_{d_0}^o} > \delta_n \forall J_{d_0}^* \right) \geq 1 - 2(d-d_0)e^{-\tau}.
\]

Hence,

\[
P(\text{`RFE finds the correct dimensions'}) \geq \prod_{i=0}^{d_0} (1 - 2(d-i)e^{-\tau}).
\]

Now for \( \tau = o(n) \) with \( \tau \to \infty \), \( P(\text{`RFE finds the correct dimensions'}) \to 1 \) as \( n \to \infty \).

Now let us prove the same assertion for the case when there is more than one correct ‘path’ from \( \mathcal{X} \) to \( \mathcal{X}^J \). Let \( J_1, \ldots, J_N \) be an enumeration of all possible such paths. Define ‘C-sets’ for a \( J_i \)
(where index \(i\) denotes the \(i^{th}\) cycle of RFE) as \(CS(J_i) := \{J_{i+1} : J_i, J_{i+1} \in \mathcal{J}_k\text{ for some }k\}\). Now,

\[
P \left( \text{‘RFE finds the correct dimensions’} \right)
\geq P \left( J_0 := J_0^o, J_1 := J_1^o \in CS(J_0^o), \ldots, J_{d_0} := J_{d_0}^o \in CS(J_{d_0-1}^o), J_{d_0+1} := \emptyset \right)
\]

\[
= P \left( J_0 := J_0^o \right) P \left( J_1 := J_1^o \in CS(J_0^o) \right) \cdots P \left( J_{d_0} := J_{d_0}^o \in CS(J_{d_0-1}^o) \right) P \left( J_{d_0+1} := \emptyset | J_{d_0}^o \right).
\]

Again as before \(P \left( J_0 := J_0^o \right) = 1\) and \(P \left( J_{d_0+1} := \emptyset | J_{d_0}^o \right) \geq 1 - 2(d - d_0)e^{-\tau}\). Now note,

\[
P \left( J_{i+1} := J_{i+1}^o \in CS(J_i^o) \mid J_i^o \right)
\]

\[
\geq P \left( \mathcal{R}_{L,D,F}^* J_{i+1}^o - \mathcal{R}_{L,D,F}^* J_i^o \leq \delta_n \forall J_{i+1}^o \in CS(J_i^o), \delta_n < \mathcal{R}_{L,D,F}^* J_{i+1}^o - \mathcal{R}_{L,D,F}^* J_i^o \forall J_{i+1}^o \notin CS(J_i^o) \right)
\]

\[
\geq 1 - \sum_{J_{i+1}^o \in CS(J_i^o)} P \left( \mathcal{R}_{L,D,F}^* J_{i+1}^o - \mathcal{R}_{L,D,F}^* J_i^o > \delta_n \right) - \sum_{J_{i+1}^o \notin CS(J_i^o)} P \left( \mathcal{R}_{L,D,F}^* J_{i+1}^o - \mathcal{R}_{L,D,F}^* J_i^o \leq \delta_n \right)
\]

\[
\geq 1 - 2|CS(J_i^o)|e^{-\tau} - 2|CS(J_i^o)|e^{-\tau} = 1 - 2(d - i)e^{-\tau},
\]

since \(|CS(J_i^o)| = |CS(J_i^o)| = d - i\). Hence again we have that

\[
P \left( \text{‘RFE finds the correct dimensions’} \right) \geq \prod_{i=0}^{d_0} \left( 1 - 2(d - i)e^{-\tau} \right).
\]

Hence for \(\tau = o(n)\) with \(\tau \to \infty\), \(P \left( \text{‘RFE finds the correct dimensions’} \right) \to 1\) as \(n \to \infty\).

(2) To prove the second part of Theorem 12 just observe that if \(J_{\text{end}}\) is the last cycle of the algorithm in RFE, then from (34) in Appendix A.7 we have that

\[
P \left( \left| \mathcal{R}_{L,P} \left( f_{D,F} J_{\text{end}} \right) - \mathcal{R}_{L,P,F}^* \right| \leq \delta_n \right) = P \left( \left| \mathcal{R}_{L,P} \left( f_{D,F} J_{\text{end}} \right) - \mathcal{R}_{L,P,F}^* \right| \leq \delta_n \right) P \left( J_{\text{end}} = J_{\text{end}} \right)
\]

\[
+ P \left( \left| \mathcal{R}_{L,P} \left( f_{D,F} J_{\text{end}} \right) - \mathcal{R}_{L,P,F}^* \right| \leq \delta_n \mid J_{\text{end}} \neq J_{\text{end}} \right) P \left( J_{\text{end}} \neq J_{\text{end}} \right)
\]

\[
\geq P \left( \left| \mathcal{R}_{L,P} \left( f_{D,F} J_{\text{end}} \right) - \mathcal{R}_{L,P,F}^* \right| \leq \delta_n \right) P \left( J_{\text{end}} = J_{\text{end}} \right)
\]

\[
\geq (1 - e^{-\tau}) \prod_{i=0}^{d_0} \left( 1 - 2(d - i)e^{-\tau} \right).
\]

So for \(\tau = o(n)\) with \(\tau \to \infty\), \(P \left( \left| \mathcal{R}_{L,P} \left( f_{D,F} J_{\text{end}} \right) - \mathcal{R}_{L,P,F}^* \right| \leq \delta_n \right) \to 1\) with \(n \to \infty\). \(\square\)

Note: Although (34) in Appendix A.7 was asserted for \(\eta_n\), it is of the same order as \(\delta_n\) and they only differ by constants. And we also have \(\eta_n < \delta_n\), so the proof for the second part of the theorem will hold true for \(\delta_n\).

**Proof.** The proof of Theorem 13 is similar to the proof of Theorem 12. Hence we omit the details and note only the needed changes. The main differences include the criterion of elimination for the algorithm, the choice of the stopping condition $\delta_n$, and the choice of $\tau$.

Note that here we use the following two assertions that, for $J_{i+1} \in CS(J_i)$, we have from the proof of (i) in Appendix A.6 that, $\lambda_n \left\| f_{D,\lambda_n,H^i} \right\|_{H_i}^2 + \tau_n \left\| f_{D,\lambda_n,H^i} \right\|_{H_i}^2 - R_{L,D} \left( f_{D,\lambda_n,H^i} \right) < \epsilon_n$, occurs with probability at least $1 - 2e^{-\tau}$ and that for $J_{i+1} \notin CS(J_i)$ being any other $J$ such that $J_i \subset J_{i+1}$ with $\|J_{i+1}\| = \|J_i\| + 1$, we also have from (27) and (28) in Appendix A.6 that $\lambda_n \left\| f_{D,\lambda_n,H^i} \right\|_{H_i}^2 + \tau_n \left\| f_{D,\lambda_n,H^i} \right\|_{H_i}^2 - R_{L,D} \left( f_{D,\lambda_n,H^i} \right) > \epsilon_0 - \epsilon_n$ occurs with probability at least $1 - 2e^{-\tau}$, for $\epsilon_n = (2c + 24\sqrt{2\tau} + 48K2n^{2p})n^{-\frac{1}{2p+1}} + 40\tau n^{-\frac{1}{2p+1}}$.

Now if we choose $\tau = o(n^{\frac{1}{2p+1}})$ with $\tau \to \infty$, then $\delta_n \equiv \epsilon_n$ satisfies the above inequalities along with the conditions that $\delta_n^{-1} = O(n^{\frac{2p}{2p+1}})$ with $\delta_n \to 0$.

Using similar steps as in the proof of Theorem 12, we have our assertion.

\[ \boxed{\text{7. Case Studies.}} \]

In this section we show the validity of our results in many practical cases of risk minimization by discussing the results in some known settings.

\[ \boxed{\text{7.1. CASE STUDY 1: Feature Elimination in Linear Regression.}} \]

In this case study we present our results for the simple setting of linear regression. This example shows that the consistency results achieved in this paper can be applied to many different situations ranging from simple to complex risk minimization problems and in some cases can substantiate known techniques that are in practice in such contexts for feature elimination. Linear regression is one of the most frequently used statistical techniques for data analysis. It is also a simple example of an empirical risk minimization problem.

In a linear regression model, we assume that the functional relationship can be expressed as $y = \langle \alpha, x \rangle + b_0$, where $\langle \alpha, x \rangle$ denotes the Euclidean inner product of vectors $\alpha$ and $x$ and $b_0$ is the bias. The prediction quality of this model can be measured by the squared-error loss function $L_{LS}$ given as $L_{LS}(x, y, f(x)) = (f(x) - y)^2$ and our goal is to find linear weights $\hat{\alpha}$ and $\hat{b}_0$ for the observed data $D$ that minimize the empirical risk. We assume that the input space $\mathcal{X} \subseteq B \subseteq \mathbb{R}^d$.

We further assume that $\mathcal{Y} \subseteq \mathbb{R}$ is a closed set. The functional space $\mathcal{F}_{lin}$ is given by $\mathcal{F}_{lin} = \{ f_{\alpha,b_0} : f_{\alpha,b_0}(x) = \langle \alpha, x \rangle + b_0, (\alpha, b_0) \in \mathbb{R}^{d+1}, \|(\alpha, b_0)\|_{\infty} \leq M, \text{ for some } M < \infty \}$. We can now observe
that the Assumptions (B1) − (B5), which are required for the consistency results, are satisfied for this problem. Note that (B1) is satisfied since the Least Squares Loss function $L_{LS}$ is convex, and as observed in SC08, $L_{LS}$ is locally Lipschitz continuous when $\mathcal{Y}$ is compact. (B3) and (B4) follow trivially under the premise that $\mathcal{X} \subseteq B \subset \mathbb{R}^d$ and $\mathcal{Y} \subset \mathbb{R}$ is a closed set and the fact that for some $M < \infty$, $\| (\alpha, b_0) \|_{\infty} \leq M$ for any function $f_{\alpha,b_0}$ within the linear functional class $\mathcal{F}_{\text{lin}}$. Since $\mathcal{F}_{\text{lin}}$ is non-empty, (B2) follows from (B3). Assumption (B5) imposes an exponential bound on the average entropy number. Many analyses have been done on covering numbers for linear function classes (see Zhang and Bartlett, 2002; Williamson, 2000) and under quite general assumptions it was proved that exponential bounds can be imposed on the $\epsilon$-entropies of such functional classes, which is actually stronger than our Assumption (B5) (Refer to Theorems 4 and 5 in Zhang and Bartlett (2002)).

Thus the RFE procedure presented in this paper translates in the linear regression case as a non-parametric backward selection method based on the value of the ‘average sum of squares of error’ or $R^2/n$. Indeed, the average empirical risk of the estimator $\hat{f}(x)$ for the sample is exactly $R^2/n$. In a non-parametric setup, under restrictive distributional assumptions on the output vector $\mathcal{Y}$, the idea of using penalized versions of $\log R^2$ like AIC, AICc or BIC are well accepted ad-hoc methodologies for model selection (and hence feature elimination), although it is not always trivial to know which penalty should be used in a given situation, or which is best in that regard. This paper produces a theoretical basis for using the non-penalized criterion $R^2/n$ as a tool for feature elimination in linear regression. Suppose we start with a set of covariates $\mathcal{X} = \{X_1, \ldots, X_d\}$ and let’s assume without loss of generality that the covariates are pre-ordered on the basis of their importance. Then Assumptions (A1) and (A2) can be interpreted as claiming the existence of an $r \in \{1, 2, \ldots, d\}$ such that the following null hypothesis is true $H_0 : \{\alpha_d = \cdots = \alpha_{r+1} = 0, \alpha_r, \ldots, \alpha_1 \neq 0\}$. So this paper establishes consistency for RFE based on the criterion $R^2/n$ and a pre-determined stopping rule in finding the correct feature space $\mathcal{X}_0 = \{X_1, \ldots, X_r\}$ under this null hypothesis $H_0$.

7.2. CASE STUDY 2: Support Vector Machines with a Gaussian RBF Kernel. Here we provide a brief review of the application of RFE in the classic SVM premise for classification using a Gaussian RBF kernel. Assume that $\mathcal{Y} = \{1, -1\}$. We want to find a function $f : \mathcal{X} \mapsto \{1, -1\}$ such that for almost every $x \in \mathcal{X}$, $P(f(x) = Y|\mathcal{X} = x) \geq 1/2$. In this case, the desired function is the Bayes
decision function $f_{L,p}^*$ with respect to the loss function $L_{BC}(x, y, f(x)) = \{y \cdot \text{sign}(f(x)) \neq 1\}$. In practice, since $L_{BC}$ is non-convex, it is usually replaced by the hinge loss function $L_{HL}(x, y, f(x)) = \max(0, 1 - yf(x))$. For SVMs with a Gaussian RBF kernel, we minimize the regularized empirical criterion $\lambda \|f\|^2 + \frac{1}{n} \sum_{i=1}^n \max(0, 1 - y_i f(x_i))$ for the observed sample $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ within the RKHS $H_\gamma(X)$ with the kernel $k_\gamma$ defined as $k_\gamma(x, y) = e^{-\|x-y\|^2_{\gamma}}$.

**Lemma 23.** For classification using support vector machines with a Gaussian RBF kernel, the RFE defined for $\delta^{-1} = O(n^{\frac{2\beta}{2d+1}})$ with $\delta \to 0$ where $\beta = \frac{\beta_0 \tau_d}{d \alpha + d \alpha + \beta_0 \tau_d}$, with $\beta_d \in (0, \infty)$ being the margin-noise exponent of the distribution $P$ on $\mathbb{R}^d \times \{-1, 1\}$ and $\tau_d \in (0, \infty]$ being the tail exponent of the marginal distribution $P_X$, is consistent in finding the correct feature space$^2$.

In order to prove Lemma 23, we need to verify Assumptions (C1) – (C6) in this setup. First note that Assumptions (C1) and (C2) are trivially satisfied since $L_{HL}$ is Lipschitz continuous (see Example 2.27 in SC08). (C3) is also satisfied since $X \in \mathbb{R}^d$ is separable and since an RKHS over separable metric spaces having a continuous kernel is separable (Lemma 4.33 of SC08), hence $H_\gamma$ is separable. It is also easy to see that $|k_\gamma(x, y)| \leq 1$ is true for all $x, y \in X$ and all $\gamma > 0$ and hence $\|k_\gamma\|_\infty \leq 1$.

From the proof of Proposition 18 (and also results in chapter 7 of SC08) we can see that Assumption (C4) can be replaced by the potentially weaker Assumption (C4*).

(C4*). Assume that for fixed $n \geq 1$, $\exists$ constants $a \geq 1$ and $p \in (0, 1)$ such that for any $J \subseteq \{1, 2, \ldots, d\}$, $\mathbb{E}_{D_X \sim P_X} e_i \left(\text{id} : H^J \mapsto L_2(D_X)\right) \leq a i^{-\frac{1}{2p}}, \quad i \geq 1$.

It is easily seen from the steps in (20) in Appendix A.5 that instead of assuming (C4) we can assume (C4*) and the results will hold. Then we see that Theorem 7.34 with Corollary 7.31 of SC08 along with the fact that $d/(d+\tau)$ is an increasing function in $d$, yields a bound as given in (C4*) with $a := \max_{d \leq \epsilon} c_{\epsilon, p} \gamma^{-\frac{(1-p)(1+\epsilon)}{2p}}$ for all $\epsilon > 0$, $d/(d+\tau) < p < 1$ and a constant $c_{\epsilon, p}$ depending only on $p$ and a given $\epsilon$. We however preferred to use (C4) instead of (C4*) in our theoretical derivations because it can be potentially weaker in many situations.

Assumption (C6) follows from results obtained in Theorem 8.18 of SC08 (see also Theorem 2.7 in Steinwart and Scovel, 2007). Note that Assumption (C6) is not required for consistency results,

$^2$For a discussion on margin-noise exponents and tail exponents of a distribution refer to Chapter 8 of SC08.
as we already have that $A_2(\lambda) \to 0$ when $\lambda \to 0$ from Lemma 5.15 of SC08. (C6) helps us to obtain explicit rates for the RFE and we show that it holds here which will help us to derive rates in this framework. Without going into explicit details, we can see from Theorem 8.18 of SC08 that the approximation error for a SVM using Gaussian RBF kernel of width $\gamma$ on $\mathbb{R}^d$ can be bounded by a function given as

$$A_2(\lambda, d, \gamma) \leq \max \{ c_{d, \tau_d}, \tilde{c}_{d, \beta_d} c_d \} \left( \lambda^{\frac{\tau_d}{d}} + \gamma \beta_d \right)$$

where $P$ is a distribution on $\mathbb{R}^d \times \{-1, 1\}$ that has margin-noise exponent $\beta_d \in (0, \infty)$ and whose marginal distribution $P_X$ has tail exponent $\tau_d \in (0, \infty]$, $c_{d, \tau_d}, \tilde{c}_{d, \beta_d} > 0$ are constants and $c_d$ is the constant occurring in equation (8.10) in SC08. So for a given pair $(\lambda, d)$ if we choose $\gamma(\lambda, d) = \lambda^{\frac{\tau_d}{d}}$ then it can be seen that $A_2(\lambda, d, \gamma(\lambda, d)) \leq \lambda^{\frac{\beta_d \tau_d}{d}}$ (where $\leq$ denotes ‘less than or equal to’ up to constants). Hence the bound in Assumption (C6) is satisfied for any $J$.

So for a sequence of SVM objective functions $\lambda_n \| f \|_{H(\lambda_n)}^2 + \frac{1}{n} \sum_{i=1}^n \max \{ 0, 1 - y_i f(x_i) \}$ defined for a sequence $\lambda_n^{-1} = o(n)$ with $\lambda_n \to 0$ the assumptions for the theoretical results on consistency of RFE are met, and thus Lemma 23 is proved.

8. Simulation Study. In this section we present a short simulation study to illustrate the use of risk-RFE for feature elimination in SVMs. Note that the use of RFE for such purposes has been in practice for well over a decade and it is a well-accepted technique in classification. The main aim of the simulation study here to evaluate our consistency results.

We consider two different data-generating mechanisms, one in the classical classification setting and the other in regression. For each of these examples we again look at three different scenarios. For the first scenario, the total number of covariates is 15 of which only 4 are important. For the second scenario, there are 30 covariates with only 7 important ones. The third scenario has 50 covariates with 3 that are important.

For the classification example we consider the hinge loss $L_{HL}$ as the surrogate loss and the SVM function is computed using the Gaussian RBF kernel $k_\gamma(x_1, x_2) = \exp\{-\frac{1}{\gamma^2} \| x_1 - x_2 \|^2 \}$. The covariates $X$ were generated uniformly on the segment $[-1, 1]$ and the output vector $Y$ was generated as $Y = \text{sign}(\omega' X)$, where $\omega$ is the vector of coefficients with the first few elements non-zero, corresponding to the important features, chosen at random from a list of coefficients.
[-1, -0.5, 0.5, 1] and the rest are zero. We initialize the original SVM function using a 5-fold cross validation on the kernel width $\gamma$ and the regularization parameter $\lambda$ and they were chosen from the set of values

\[(16) \quad \left(\frac{2}{n\lambda}, \gamma\right) = (0.01 \times 10^i, j) \quad i = \{0, 1, 2, 3, 4\}, \quad j = \{1, 2, 3, 4\}\]

where $n$ is the sample size for the given setting.

In the second case we used an SVR function with a linear kernel $k(x_1, x_2) = \langle x_1, x_2 \rangle$ to treat the regression setting. The loss function we considered is the $\epsilon$-insensitive Loss $L_{\epsilon}(x, y, f(x)) = \max\{0, |y - f(x)| - \epsilon\}$ with $\epsilon = 0.1$. Covariates are generated as before while $Y$ is now generated as $Y = \omega'X + \frac{1}{3}N_{\dim(X)}(0, 1)$. As before we initialize with a 5-fold Cross Validation on $\lambda$.

We repeat the process for different sample sizes $n = \{100, 200, 400\}$. We also repeat the simula-

---

### Table 1. Accuracy of Risk-RFE

<table>
<thead>
<tr>
<th>SVM with Gaussian Kernel</th>
<th>$d = 15, d_0 = 4$</th>
<th>$d = 30, d_0 = 7$</th>
<th>$d = 50, d_0 = 3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prop of times no errors (a)</td>
<td>0.97</td>
<td>0.62</td>
<td>0.95</td>
</tr>
<tr>
<td>Prop of times 1 error (b)</td>
<td>0.03</td>
<td>0.34</td>
<td>0.05</td>
</tr>
<tr>
<td>Prop of times &gt; 1 error (c)</td>
<td>0</td>
<td>0.04</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SVR with Linear Kernel</th>
<th>$d = 15, d_0 = 4$</th>
<th>$d = 30, d_0 = 7$</th>
<th>$d = 50, d_0 = 3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prop of times no errors (a)</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Prop of times 1 error (b)</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Prop of times &gt; 1 error (c)</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

---

![Reverse Scree Graph for one run of the simulations for (a) SVM with Gaussian Kernel (b) SVR with Linear Kernel with $d = 30, d_0 = 7$](image)
Fig 2. Linear-Quadratic mixture change point analysis for (a) SVM with Gaussian Kernel for comparable cross validation values of $\lambda$ and kernel width $\gamma$ and (b) SVR with Linear Kernel for comparable cross validation values of $\lambda$, with $d = 30$, $d_0 = 7$ for varying sample sizes. The bold dots represent the estimated change points.

Fig 3. Linear-Quadratic mixture change point analysis for (a) SVM with Gaussian Kernel for comparable cross validation values of $\lambda$ and kernel width $\gamma$ and (b) SVR with Linear Kernel for comparable cross validation values of $\lambda$, with $d = 50$, $d_0 = 3$ for varying sample sizes. The bold dots represent the estimated change points.

tions 100 times each to note down the proportion of times the Risk-RFE made no errors (a), made only one error (b) or made more than 1 error (c) (See Table 8), where a mistake is made if the rank of any non-important feature is found to be higher than that of any important one. The entire methodology was implemented in the MATLAB environment. For the implementation we used the SPIDER library for MATLAB\(^3\), which already has a feature elimination algorithm based on RFE and we modified it accordingly to suit our criterion for reduction. The codes for the algorithm and the simulations are given in Supplement A.

\(^3\)The Spider library for Matlab can be downloaded from [http://www.kyb.tuebingen.mpg.de/bs/people/spider/](http://www.kyb.tuebingen.mpg.de/bs/people/spider/)
One important question we inevitably face in feature elimination is when to stop. Note that our theoretical results suggest the existence of a gap $\epsilon_0$ and our results show that asymptotically the difference in the empirical versions of the objective functions exceed it whenever we move beyond the correct dimension. Practically it is almost impossible to characterize this gap for a given setting, but the existence of this gap can be observed from the values of the objective function at each stage of the algorithm. One idea that can be implemented is that of a ‘reverse Scree graph’ (See section on Scree graphs in chapter 6 of Jolliffe (2002)). Implementation of the Scree graph is a well-formulated idea in choosing the correct number of Principal Components in PCA and that same idea can be applied here as well. We plot the values of the objective function
\[
\inf_{i \in \mathbb{Z} \setminus J} \lambda \left\| f_{D,\lambda,H,J \cup \{i\}} \right\|_{H,J \cup \{i\}}^2 + R_{L,D} \left( f_{D,\lambda,H,J \cup \{i\}} \right)
\]
at each run of the algorithm in a graph. Figure 1 justifies such an argument.

For a further exploratory analysis of this gap and to characterize the number of features to be eliminated, we tried some ad-hoc model diagnostic tools. From a heuristic standpoint, the phenomenon captured in Figure 1 seems to suggest that if we fit a regression model to the observed objective function values in the scree plot, we will expect a change in the slope of the regression line right after we start eliminating significant covariates because of the aforementioned gap. One plausible way to analyze this gap is to fit a change point regression model of the observed values on the number of cycles of RFE and to infer that the estimated change point is the ad-hoc stopping rule, so as to eliminate all features ranked below that point. For the asymptotic belief that the change in the objective function is negligible to the left of the change point, we fit a linear trend there. However to the right of the change point, these changes might show non-linear trends, and hence we tried linear and quadratic trends to model that. The quadratic trend seemed to work better. Some plots (see Figures 2, 3) are given here to show our analysis where we show the mixture of linear-quadratic fits.

So heuristically it is possible to justify the choice of the correct dimensions based on a reverse Scree graph. Otherwise it is a user-determined choice for the gap size to determine how many dimensions are required in a specific setting.

9. Discussion. We proposed an algorithm for feature elimination in empirical risk minimization and support vector machines. We studied the theoretical properties of the method, discussed
the necessary assumptions, and showed that it is universally consistent in finding the correct feature space under these assumptions. We provided case studies of a few of the many different scenarios where this method can be used. Finally, we give a short simulation study to illustrate the method and discuss a practical method for choosing the correct subset of features.

Note that Lemma 21(iii) and Lemma 22(iii) establish the existence of a gap in the rate of change of the objective function at the point where our feature elimination method begins removing essential features of the learning problem. This motivated us to use a scree plot of the values of the objective function at each cycle, and indeed our simulation results support our approach by visually exhibiting this gap. Moreover, the graphical interpretation of the scree plot motivated the use of change point regression to select the correct feature space. It would be interesting to conduct a more detailed and formal analysis of this gap in real life settings to facilitate more efficient, automated practical solutions.

As far as our knowledge goes, not much analysis have been done on the properties of variable selection algorithms under such general assumptions on the probability generating mechanisms of the input space, especially in support vector machines. So the results generated in this paper can act as a good starting point for similar analyses in other settings. One useful extension would be to study the properties of RFE under the high dimensional framework of a ‘large $p$ small $n$ problem’. It would also be interesting to analyze RFE for other settings, including censored support vector regression (See Goldberg and Kosorok (2013)) or other machine learning problems, including reinforcement learning or other penalized risk minimization problems.

APPENDIX A: PROOFS

A.1. Proof of Lemma 5.

**Proof.** The direction $L^J_\infty(\mathcal{X}^J) \subseteq L_\infty(\mathcal{X}^J)$ is obvious since co-ordinate projection maps are continuous. To show that $L^J_\infty(\mathcal{X}^J) \supseteq L_\infty(\mathcal{X}^J)$ let us take $g \in L_\infty(\mathcal{X}^J)$. Then $g : \mathcal{X}^J \mapsto \mathbb{R}$ is measurable with $\|g\|_\infty < \infty$. Extend $g$ to $\bar{g}$ to include the whole domain $\mathcal{X}$ by defining $\bar{g}(x) = g (\pi^J(x))$. Since $\bar{g}$ is measurable with $\|\bar{g}\|_\infty = \|g\|_\infty$, we have that $\bar{g} \in L_\infty(\mathcal{X})$ and $\bar{g} \circ \pi^J = \bar{g}$, so $g = \bar{g} |_{\mathcal{X}^J} \in L^J_\infty(\mathcal{X}^J)$. 

Hence \( F \) is dense in \( L_\infty(\mathcal{X}) \).

(2) Since \( F \) is compact, for any \( \epsilon > 0 \), \( \exists \{f_n\}_{n=1}^{N_\epsilon} \subset F \) such that \( F \subset \bigcup_{n=1}^{N_\epsilon} B_{\|\cdot\|_\infty}(f_n, \epsilon) \) (where \( B_{\|\cdot\|_\infty}(f_n, \epsilon) \) is a \( \|\cdot\|_\infty \) ball of radius \( \epsilon \) with center \( f_n \)). We now fix \( f \in F^J \) and note that \( \exists \) an equivalent class of functions \( \{g^J\} \) in \( F \) such that for any two functions \( g_1^J \) and \( g_2^J \in \{g^J\} \) we have that \( g_1^J \sim g_2^J \) in the sense that \( g_1^J \circ \pi^Jc = g_2^J \circ \pi^Jc = f \). Fix one such \( g^J \in \{g^J\} \). Since \( g^J \in F, \exists f_i \in \{f_n\}_{n=1}^{N_\epsilon} \) such that \( d(f_i, g^J) < \epsilon \), that is,

\[
\sup_{x \in \mathcal{X}} |f_i(x) - g^J(x)| < \epsilon \quad \Rightarrow \quad \sup_{x \in \pi^Jc(\mathcal{X})} |f_i(x) - g^J(x)| < \epsilon
\]

\[
\Rightarrow \quad \sup_{x \in \mathcal{X}} |f_i(\pi^Jc(x)) - g^J(\pi^Jc(x))| < \epsilon \quad \Rightarrow \quad \sup_{x \in \mathcal{X}} |f_i^J(x) - f(x)| < \epsilon \quad (\because g^J(\pi^Jc(x)) = f(x))
\]

\[
\Rightarrow \quad \{f_n\}_{n=1}^{N_\epsilon} \text{ forms a finite } \epsilon \text{-cover for the set } F^J.
\]

Hence \( F^J \) is compact.

(3) To see (3), note that if \( f_1, \ldots, f_{2^n-1} \) is an \( \epsilon \)-net of \( F \), then for any \( f \in F \), we have \( i \in \{1, \ldots, 2^{n-1}\} \) such that \( \|f - f_i\|_\infty < \epsilon \). Then,

\[
\|f \circ \pi^Jc - f_i \circ \pi^Jc\|_\infty = \sup_{x \in \mathcal{X}} |f \circ \pi^Jc(x) - f_i \circ \pi^Jc(x)| = \sup_{x \in \mathcal{X}^J} |f(x) - f_i(x)| \leq \sup_{x \in \mathcal{X}} |f(x) - f_i(x)| = \|f - f_i\|_\infty < \epsilon.
\]

Hence \( f_1 \circ \pi^Jc, \ldots, f_{2^n-1} \circ \pi^Jc \) is an \( \epsilon \)-net of \( F^J \). \( \square \)

PROOF. Note that if we define \( g_f := L \circ f - E_P(L \circ f) \), then \( G = \{g_f : f \in \mathcal{F}\} \) is a separable Carathéodory set (for a discussion on Carathéodory families of maps, refer to Definition 7.4 in SC08). To see this, first note that \( \|g_f\|_\infty \leq \sup_{(x,y) \in \mathcal{X} \times \mathcal{Y}} |L \circ f - E_P(L \circ f)| \leq 2B \) for \( B \) defined in the statement of the Lemma. Also by assumption, \( \|\cdot\| \) dominates the pointwise convergence of functions (so \( f_n \rightarrow f \) in \( \|\cdot\| \Rightarrow f_n \rightarrow f \) pointwise). Then the fact that \( L \) is locally-Lipschitz continuous coupled with Lebesgue’s Dominated Convergence Theorem (since \( \|L \circ f\|_\infty \leq B \)) gives us the above assertion.

Now note that \( E_P(g_f) = 0 \) and \( E_P g_f^2 \leq (2B)^2 = 4B^2 \) for \( B \) as before, so we can apply the Talagrand’s Inequality given in Theorem 7.5 of SC08 on \( G \) defined as \( G : Z^n \equiv (\mathcal{X} \times \mathcal{Y})^n \rightarrow \mathbb{R} \) such that

\[
G(z_1, \ldots, z_n) := \sup_{g_f \in G} \left| \frac{1}{n} \sum_{j=1}^{n} g_f(z_j) \right| = \sup_{f \in \mathcal{F}} |\mathcal{R}_{L,D}(f) - \mathcal{R}_{L,P}(f)|,
\]

and hence, for \( \gamma = 1 \) and for all \( \tau > 0 \), we have

\[
P^n \left( \left\{ z \in Z^n : G(z) \geq 2E_P^n(G) + 2B \sqrt{\frac{2\tau}{n} + \frac{10B\tau}{3n}} \right\} \right) \leq e^{-\tau}.
\]

So now we need to bound the term \( E_P^n(G) := E_P^n \left\{ \sup_{f \in \mathcal{F}} |\mathcal{R}_{L,D}(f) - \mathcal{R}_{L,P}(f)| \right\} \). We have for all \( n \geq 1 \),

\[
E_{D \sim P^n} \left\{ \sup_{f \in \mathcal{F}} |\mathcal{R}_{L,D}(f) - \mathcal{R}_{L,P}(f)| \right\} = E_{D \sim P^n} \sup_{h \in \mathcal{H}} |E_P h - E_D h| \leq 2E_{D \sim P^n} \text{Rad}_D(\mathcal{H}, n),
\]

where \( \text{Rad}_D(\mathcal{H}, n) \) is the \( n \)-th empirical Rademacher average of \( \mathcal{H} \) for \( D := \{z_1, \ldots, z_n\} \in Z^n \) with respect to the Rademacher sequence \( \{\varepsilon_1, \ldots, \varepsilon_n\} \) and the distribution \( \nu \), which is given by \( \text{Rad}_D(\mathcal{H}, n) = E_\nu \sup_{h \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i h(z_i) \right| \). So we see now that it suffices to bound \( E_{D \sim P^n} \text{Rad}_D(\mathcal{H}, n) \).

For that we use theorem 7.16 of SC08, but before that note from (C4) that we have that for fixed \( n \geq 1 \), that \( \exists \) constants \( a \geq 1 \) and \( p \in (0, 1) \) such that

\[
E_{D \sim P^n} e_i(\mathcal{F} \mathcal{L}_\infty(D\mathcal{X})) \leq a^{- \frac{1}{2p}}, \quad i \geq 1.
\]
First observe that $\mathcal{H} \subset \mathcal{L}_2(P)$. Now since Lipschitz continuity of $L$ gives us that $|L(x, y, f_1(x)) - L(x, y, f_2(x))|^2 \leq c_L(C)^2 |f_1(x) - f_2(x)|^2$, it is easy to see that

$$e_i(\mathcal{H}, \| \cdot \|_{L_2(P)}) \leq c_L(C)e_i(\mathcal{F}, \| \cdot \|_{L_2(P_{\mathcal{X}})}).$$

Hence we have

$$E_{D \sim P^n} \left( e_i(\mathcal{H}, \| \cdot \|_{L_2(D)}) \right) \leq c_L(C)E_{D \sim P^n} \left( e_i(\mathcal{F}, \| \cdot \|_{L_2(D)}) \right) \leq c_L(C)E_{D \sim P^n} \left( e_i(\mathcal{F}, \| \cdot \|_{L_\infty(D)}) \right) \leq c_L(C)\bar{a}^{-\frac{1}{p}}.$$

Now noting that $\|h_i\|_\infty \leq B$ and $E_p h_j^2 \leq B^2$ for $B$ defined as before, the conditions of Theorem 7.16 of SC08 are satisfied with $\bar{a} = c_L(C)a$ and hence we have,

$$E_{D \sim P^n} \text{Rad}_D(\mathcal{H}, n) \leq \max \left\{ C_1(p)\bar{a}^p B^{1-p} n^{-\frac{1}{2}}, C_2(p)\bar{a}^{\frac{2p}{1+p}} B^{\frac{1-p}{1+p}} n^{-\frac{1}{1+p}} \right\}$$

for constants $C_1(p), C_2(p)$ depending only on $p$. Hence we finally have, that with probability $\geq 1 - e^{-\tau},$

$$\sup_{f \in \mathcal{F}} |\mathcal{R}_{L,P}(f) - \mathcal{R}_{L,D}(f)| \leq 2B \sqrt{\frac{2\tau}{n}} + \frac{10B\tau}{3n} + 4 \max \left\{ C_1(p)c_L(C)^p a^p B^{1-p} n^{-\frac{1}{2}}, C_2(p)c_L(C)^{\frac{2p}{1+p}} a^{\frac{2p}{1+p}} B^{\frac{1-p}{1+p}} n^{-\frac{1}{1+p}} \right\}.$$

\[\Box\]


PROOF. For $a^{2p} > n$, note that $|\mathcal{R}_{L,D}(f_{D,F_{J_2}} - \mathcal{R}_{L,D}(f_{D,F_{J_1}})| \leq \mathcal{R}_{L,D}(f_{D,F_{J_2}}) + \mathcal{R}_{L,D}(f_{D,F_{J_1}}) \leq 2B \leq 24K_1 \leq 24K_1 \left( \frac{2p}{n} \right)^{1/2}$ for $K_1 \geq B/4$. Hence we assume $a^{2p} \leq n$. Now note that for $J_1, J_2 \in \tilde{J}$ such that $J_1 \subseteq J_2 \subseteq J_*$, we have,

$$|\mathcal{R}_{L,D}(f_{D,F_{J_2}}) - \mathcal{R}_{L,D}(f_{D,F_{J_1}})| \leq |\mathcal{R}_{L,D}(f_{D,F_{J_2}}) - \mathcal{R}_{L,P}(f_{D,F_{J_2}})| + |\mathcal{R}_{L,P}(f_{D,F_{J_2}}) - \mathcal{R}_{L,P,F_{J_2}}| + |\mathcal{R}_{L,P,F_{J_1}} - \mathcal{R}_{L,P}(f_{D,F_{J_1}})| + |\mathcal{R}_{L,P}(f_{D,F_{J_1}}) - \mathcal{R}_{L,D}(f_{D,F_{J_1}})|$$

$$\leq A_n^1 + B_n^1 + B_n^2 + A_n^2$$

since from Assumption (A1) we have that $\mathcal{R}_{L,P}(f_{P,F_{J_1}}) = \mathcal{R}_{L,P}(f_{P,F_{J_2}})$ for $J_1 \subseteq J_2 \subseteq J_*$. For $A_n^1$ and $A_n^2$ we have, $|\mathcal{R}_{L,D}(f_{D,F_{J_2}}) - \mathcal{R}_{L,P}(f_{D,F_{J_2}})| \leq \sup_{f \in F_{J_2}} |\mathcal{R}_{L,P}(f) - \mathcal{R}_{L,D}(f)|$ and $|\mathcal{R}_{L,D}(f_{D,F_{J_1}}) - \mathcal{R}_{L,P}(f_{D,F_{J_1}})| \leq \sup_{f \in F_{J_1}} |\mathcal{R}_{L,P}(f) - \mathcal{R}_{L,D}(f)|$. 

Hosted by The Berkeley Electronic Press
Now see from Lemma 6 that the conditions of Lemma 14 are satisfied for $B$

\[ |\mathcal{R}_{L,P}(f_{D,F,t_2}) - \mathcal{R}_{L,P}(f_{P,F,t_2})| = \mathcal{R}_{L,P}(f_{D,F,t_2}) - \mathcal{R}_{L,P,F,t_2}^* \leq 2 \sup_{f \in F,t_2} |\mathcal{R}_{L,P}(f) - \mathcal{R}_{L,D}(f)|, \]

\[ |\mathcal{R}_{L,P}(f_{D,F,t_1}) - \mathcal{R}_{L,P}(f_{P,F,t_1})| = \mathcal{R}_{L,P}(f_{D,F,t_1}) - \mathcal{R}_{L,P,F,t_1}^* \leq 2 \sup_{f \in F,t_1} |\mathcal{R}_{L,P}(f) - \mathcal{R}_{L,D}(f)|. \]

Now see from Lemma 6 that the conditions of Lemma 14 are satisfied for $\|\cdot\|_F := \|\cdot\|_{\infty}, C := M$ and $B := B$ for each of the functional classes $F$. Also since $a^{2p} \leq n$, we have $\left(\frac{a^{2p}}{n}\right)^{1/2} \geq \left(\frac{a^{2p}}{n}\right)^{1/(p+1)}$ for $p \in (0, 1)$. Hence we have our assertion.

\[ \square \]


**Proof.** First note that since $B \geq 1$ and $K \geq B^p/4$, we have $24KB^{1-p} \geq 6B > 2$. Now if $a^{2p} > \lambda^p n$, the inequality trivially follows from the fact that

\[ \lambda \|f_{D,\lambda,H,t_2}\|_{H,t_2}^2 + \mathcal{R}_{L,D}(f_{D,\lambda,H,t_2}) - \lambda \|f_{D,\lambda,H,t_1}\|_{H,t_1}^2 - \mathcal{R}_{L,D}(f_{D,\lambda,H,t_1}) \]

\[ \leq \lambda \|f_{D,\lambda,H,t_1}\|_{H,t_2}^2 + \mathcal{R}_{L,D}(f_{D,\lambda,H,t_2}) + \lambda \|f_{D,\lambda,H,t_1}\|_{H,t_1}^2 + \mathcal{R}_{L,D}(f_{D,\lambda,H,t_1}) \]

\[ \leq 2\mathcal{R}_{L,D}(0) \leq 24KB^{1-p} \left(\frac{a^{2p}}{\lambda^p n}\right)^{\frac{1}{2}}, \]

since $\mathcal{R}_{L,D}(0) \leq 1$. Hence we assume from here on that $a^{2p} \leq \lambda^p n$. Now observe that since $H$ is separable, from Lemma 8 we have that the $H^J$s are also separable. Hence from Lemma 6.23 of SC08 we have that the SVMs produced by these RKHSs are measurable.

Now note that $L(x, y, 0) \leq 1 \Rightarrow$ for any distribution $Q$ on $\mathcal{X} \times \mathcal{Y}$, we have that $\mathcal{R}_{L,Q}(0) \leq 1$. Since, \[ \inf_{f \in H^J} \lambda \|f\|_{H^J}^2 + \mathcal{R}_{L,Q}(f) \leq \mathcal{R}_{L,Q}(0), \]

we have that $\|f_{Q,\lambda,H^J}\|_{H^J} \leq \sqrt{\frac{\mathcal{R}_{L,Q}(0)}{\lambda}}$. Now since by Lemma 4.23 of SC08 $\|f\|_{\infty} \leq \|k\|_{\infty} \|f\|_{H^J}$ for all $f \in H^J$, we have that $\|f_{Q,\lambda,H^J}\|_{\infty} \leq \|f_{Q,\lambda,H^J}\|_{H^J} \leq \lambda^{-1/2}$. So, consequently, for every distribution $Q$ on $\mathcal{X} \times \mathcal{Y}$, we have

\[ (22) \quad |\mathcal{R}_{L,P}(f_{Q,H^J}) - \mathcal{R}_{L,D}(f_{Q,H^J})| \leq \sup_{f \in H^J} \|f\|_{H^J} \leq \lambda^{-1/2}. \]

Now,

\[ \lambda \|f_{D,\lambda,H,t_2}\|_{H,t_2}^2 + \mathcal{R}_{L,D}(f_{D,\lambda,H,t_2}) - \lambda \|f_{D,\lambda,H,t_1}\|_{H,t_1}^2 - \mathcal{R}_{L,D}(f_{D,\lambda,H,t_1}) \]

\[ \leq \lambda \|f_{D,\lambda,H,t_1}\|_{H,t_2}^2 + \mathcal{R}_{L,D}(f_{D,\lambda,H,t_2}) - \lambda \|f_{D,\lambda,H,t_1}\|_{H,t_1}^2 + \mathcal{R}_{L,D}(f_{D,\lambda,H,t_1}) \]

\[ + \lambda \|f_{D,\lambda,H,t_2}\|_{H,t_1}^2 + \mathcal{R}_{L,P}(f_{D,\lambda,H,t_2}) - \lambda \|f_{D,\lambda,H,t_1}\|_{H,t_1}^2 + \mathcal{R}_{L,P}(f_{D,\lambda,H,t_1}) \]

\[ + \lambda \|f_{D,\lambda,H,t_2}\|_{H,t_1}^2 + \mathcal{R}_{L,P}(f_{D,\lambda,H,t_2}) - \lambda \|f_{D,\lambda,H,t_1}\|_{H,t_1}^2 + \mathcal{R}_{L,P}(f_{D,\lambda,H,t_1}) \],
since from (A1), \( \mathcal{R}^*_{L,P,H^1} = \mathcal{R}^*_{L,P,H^2} = \mathcal{R}^*_{L,P,H} \). Noting that
\[
\lambda \| f_{D,\lambda,H^2} \|_{H^2}^2 + \mathcal{R}_{L,P} (f_{D,\lambda,H^2}) - \mathcal{R}^*_{L,P,H^2} \geq 0,
\]
we have from (6.18) of SC08 that
\[
\left| \lambda \| f_{D,\lambda,H^2} \|_{H^2}^2 + \mathcal{R}_{L,P} (f_{D,\lambda,H^2}) - \mathcal{R}^*_{L,P,H^2} \right| \\
\leq A_2^J (\lambda) + \mathcal{R}_{L,P} (f_{D,\lambda,H^2}) - \mathcal{R}_{L,D} (f_{D,\lambda,H^2}) + \mathcal{R}_{L,D} (f_{P,\lambda,H^2}) - \mathcal{R}_{L,P} (f_{P,\lambda,H^2}) \\
\leq A_2^J (\lambda) + 2 \sup_{\|f\|_{H^2} \leq \lambda^{-1/2}} |\mathcal{R}_{L,P}(f) - \mathcal{R}_{L,D}(f)|.
\]
(23)

From (22) and (23) and the fact that \( J_1, J_2 \in \bar{J} \) such that \( J_1 \subseteq J_2 \subseteq J_\tau \), we have that
\[
\left| \lambda \| f_{D,\lambda,H^2} \|_{H^2}^2 + \mathcal{R}_{L,D} (f_{D,\lambda,H^2}) - \lambda \| f_{D,\lambda,H^1} \|_{H^1}^2 - \mathcal{R}_{L,D} (f_{D,\lambda,H^1}) \right| \\
\leq A_2^J (\lambda) + A_2^J (\lambda) + 3 \sup_{\|f\|_{H^1} \leq \lambda^{-1/2}} |\mathcal{R}_{L,P}(f) - \mathcal{R}_{L,D}(f)| + 3 \sup_{\|f\|_{H^2} \leq \lambda^{-1/2}} |\mathcal{R}_{L,P}(f) - \mathcal{R}_{L,D}(f)|.
\]

First note that for \( f \in \lambda^{-1/2} \mathcal{B}_{H^J} \) and \( B := c_L (\lambda^{-1/2}) \lambda^{-1/2} + 1 \), we have \( |L(x,y,f(x))| \leq |L(x,y,f(x)) - L(x,y,0)| + L(x,y,0) \leq B \) for all \( (x,y) \in \mathcal{X} \times \mathcal{Y} \). Also note that Assumption (C4) implies that \( E_{D_{\mathcal{X}} \sim P_{\mathcal{X}}}(e_{i}(\lambda^{-1/2} \mathcal{B}_{H^l}, \| : \|_{L_{\infty}(D_{\mathcal{X}})})) \leq \lambda^{-1/2} a i^{-\frac{1}{2p}} \).

Now note from Lemma 8 that the conditions of Lemma 14 are satisfied for \( \mathcal{F} := \lambda^{-1/2} \mathcal{B}_{H^J} \), \( \| \cdot \| := \| \cdot \|_{H^J} \), \( C := \lambda^{-1/2} \) and \( B := c_L (\lambda^{-1/2}) \lambda^{-1/2} + 1 \) for each of the RKHS classes \( H^J \). Also since \( a^{2p} \leq \lambda^p n \) and \( B \geq 1 \), we have \( \left( \frac{a^{2p}}{n} \right)^{1/2} \geq \left( \frac{a^{2p}}{n} \right)^{1/(p+1)} \) and \( B^{1-p} \geq B^{\frac{1-p}{1+p}} \) for \( p \in (0,1) \). Hence we have our assertion.

**A.6. Proof of Lemma 22.**

**Proof.** (i) Fixing a \( \lambda \in [0,1] \), we have that \( B := c_L (\lambda^{-1/2}) \lambda^{-1/2} + 1 \leq 2 \lambda^{-1/2} \). Now since \( |X| \leq x \Rightarrow X \leq x \) for any \( x \geq 0 \), we see from Proposition 18 that
\[
\lambda \| f_{D,\lambda,H^2} \|_{H^2}^2 + \mathcal{R}_{L,D} (f_{D,\lambda,H^2}) - \lambda \| f_{D,\lambda,H^1} \|_{H^1}^2 - \mathcal{R}_{L,D} (f_{D,\lambda,H^1}) \\
< A_2^J (\lambda) + A_2^J (\lambda) + 24 \lambda^{-1/2} \sqrt{\frac{2\tau}{n}} + 40 \lambda^{-1/2} \frac{\tau}{n} + 48 K_2 \lambda^{-\frac{p-1}{2}} \left( \frac{a^{2p}}{n} \right)^{\frac{1}{2}} \\
= A_2^J (\lambda) + A_2^J (\lambda) + 24 \sqrt{2\tau} (\lambda n)^{-\frac{1}{2}} + 40 \tau (\lambda^{\frac{3}{2}} n)^{-1} + 48 K_2 a^{2p} (\lambda n)^{-\frac{1}{2}}
\]
with probability at least \( 1 - 2e^{-\tau} \). Also from Corollary 19, for \( J \in \bar{J} \) similarly, we have
\[
\left| \lambda \| f_{D,\lambda,H^J} \|_{H^J}^2 + \mathcal{R}_{L,D} (f_{D,\lambda,H^J}) - \mathcal{R}^*_{L,P,H} \right| \\
< A_2^J (\lambda) + 12 \sqrt{2\tau} (\lambda n)^{-\frac{1}{2}} + 20 \tau (\lambda^{\frac{3}{2}} n)^{-1} + 24 K_2 a^{2p} (\lambda n)^{-\frac{1}{2}}
\]
(25)
with probability at least $1 - e^{-\tau}$. Then under Assumption (C5) and Lemma 5.15 along with (5.32) of SC08 we obtain that the right hand side of the above inequality converges to 0. So the denseness assumption of the RKHSs additionally gives us the universal consistency of our feature elimination algorithm. To establish the convergence rate of our algorithm we further assume (C6) that there exists $c > 0$ and $\beta \in (0, 1]$ such that $A^J \leq c\lambda^J$ for any $J$ and for all $\lambda \geq 0$. Then it can be seen that asymptotically the best choice for $\lambda_n$ in (24) or (25) is a sequence that behaves like $n^{-\frac{\beta}{2(3\beta + 1)}}$ and then the inequalities in (24) and (25) are satisfied with the l.h.s. replaced by $\epsilon_n$ and $\epsilon_n/2$ respectively, where $\epsilon_n$ is given by $(2c + 24\sqrt{2\tau} + 48\xi_2a^{2p})n^{-\frac{\beta}{2(3\beta + 1)}} + 40\tau n^{-\frac{4\beta + 1}{2(3\beta + 1)}}$. This proves (i) for $\{\epsilon_n\}$ for a suitable choice of $\tau$.

(ii) Observe from Corollary 19 along with Assumptions (C5), (C6) and steps in proof of (i) given above that,

$$\left| \lambda_n \left\| f_{D,\lambda_n,H^J} \right\|_{H^J}^2 + R_{L,D}\left( f_{D,\lambda_n,H^J} \right) - R^*_{L,P,H^J} \right| < \epsilon_n/2$$

occurs with $P^n$ probability greater than $1 - e^{-\tau}$ for any $J \subset \{1, 2, \ldots, d\}$ where $\epsilon_n$ is given as before.

Also note that from Assumption (A2) we have that $R^*_{L,P,H^J} - \epsilon_0 \geq R^*_{L,P,H^J_2} = R^*_{L,P,H^J_1}$. So for $H^J_2$ we have,

$$P^n \left( D \in (X \times Y)^n : \lambda_n \left\| f_{D,\lambda_n,H^J_2} \right\|_{H^J_2}^2 + R_{L,D}\left( f_{D,\lambda_n,H^J_2} \right) - R^*_{L,P,H^J_2} \right| < \delta_n \right) > 1 - e^{-\tau}$$

(27)

$$\Rightarrow P^n \left( D \in (X \times Y)^n : \lambda_n \left\| f_{D,\lambda_n,H^J_2} \right\|_{H^J_2}^2 + R_{L,D}\left( f_{D,\lambda_n,H^J_2} \right) + \delta_n > R^*_{L,P,H^J_2} \right) > 1 - e^{-\tau},$$

and for $H^J_1$ we have

$$P^n \left( D \in (X \times Y)^n : \lambda_n \left\| f_{D,\lambda_n,H^J_1} \right\|_{H^J_1}^2 + R_{L,D}\left( f_{D,\lambda_n,H^J_1} \right) - R^*_{L,P,H^J_1} \right| < \delta_n \right) > 1 - e^{-\tau}$$

(28)

$$\Rightarrow P^n \left( D \in (X \times Y)^n : \lambda_n \left\| f_{D,\lambda_n,H^J_1} \right\|_{H^J_1}^2 + R_{L,D}\left( f_{D,\lambda_n,H^J_1} \right) < R^*_{L,P,H^J_1} + \delta_n \right) > 1 - e^{-\tau},$$

Then (27) and (28) from above jointly imply that

$$\lambda_n \left\| f_{D,\lambda_n,H^J_2} \right\|_{H^J_2}^2 + R_{L,D}\left( f_{D,\lambda_n,H^J_2} \right) - \lambda_n \left\| f_{D,\lambda_n,H^J_1} \right\|_{H^J_1}^2 - R_{L,D}\left( f_{D,\lambda_n,H^J_1} \right) > \epsilon_0 - 2\delta_n$$

(29)

with $P^n$ probability greater than $1 - 2e^{-\tau}$. 

http://biostats.bepress.com/uncbiostat/art37
Also it is easy to see that since $\epsilon_n \to 0$ with $n \to \infty$, the gap $\tilde{\epsilon}_n = \epsilon_0 - \epsilon_n \to \epsilon_0 > 0$.

(iii) From Assumption (A1), Corollary 20, additional Assumptions (C5), (C6) and steps in the proof of (i) given above, the ‘if’ condition of (iii) follows since for any $\epsilon > 0$, $\tau > 0$ and $n \geq 1$ we have,

$$P^n \left( D \in (\mathcal{X} \times \mathcal{Y})^n : \lambda_n \left\| f_{D,\lambda_n} \right\|_{H^J}^2 + \mathcal{R}_{L,P} \left( f_{D,\lambda_n} - \mathcal{R}_L^{*} \right) < \eta_n \right) > 1 - e^{-\tau},$$

where $\eta_n = (c + 8\sqrt{2\tau} + 16K_2a^p)n^{-\frac{\beta}{\beta + 1}} + 40/3\tau n^{-\frac{4\beta + 1}{4\beta + 1}}$.

Now for $J_1 \in \tilde{\mathcal{J}}$ and $J_2 \notin \tilde{\mathcal{J}}$ we have $\mathcal{R}_{L,P,F,J_2} - \epsilon_0 \geq \mathcal{R}_{L,P,F,J_1} = \mathcal{R}_{L,P,F,J_1}$ and hence the ‘only if’ condition of (iii) also follows since

$$P^n \left( D \in (\mathcal{X} \times \mathcal{Y})^n : \lambda_n \left\| f_{D,\lambda_n} \right\|_{H^J}^2 + \mathcal{R}_{L,P} \left( f_{D,\lambda_n} - \mathcal{R}_L^{*} \right) > \epsilon_0 - \delta_n \right) > 1 - e^{-\tau}.$$

Now since $\eta_n \to 0$ with $n \to \infty$, the gap $\tilde{\epsilon}_n = \epsilon_0 - \eta_n \to \epsilon_0 > 0$. \hfill \Box


Proof. (i) Note that Proposition 15 implies that we have with probability at least $1 - 2e^{-\tau}$,

$$\mathcal{R}_{L,D} \left( f_{D,F,J_2} \right) < \mathcal{R}_{L,D} \left( f_{D,F,J_1} \right) + 12B\sqrt{\frac{2\tau}{n}} + \frac{20B\tau}{n} + 24K_1 \left( \frac{a^p}{n} \right)^{\frac{1}{2}}.$$

This trivially establishes (i) for $\epsilon_n = 12B\sqrt{2\tau n^{-1/2} + 20B\tau n^{-1} + 24K_1 a^p n^{-1/2}}$ for sufficiently large $\tau$.

(ii) Corollary 16 gives us that $|\mathcal{R}_{L,D} \left( f_{D,F,J} \right) - \mathcal{R}_{L,D}^{*} | < \epsilon_n/2$ with probability at least $1 - e^{-\tau}$ for any $J \subset \{1, 2, \ldots, d\}$, where $\epsilon_n$ is as defined in (i).

Now, following similarly as the steps given in the proof of Lemma 22 in Appendix A.6, we obtain that $P^n \left( D \in (\mathcal{X} \times \mathcal{Y})^n : \mathcal{R}_{L,D} \left( f_{D,F,J_2} \right) - \mathcal{R}_{L,D} \left( f_{D,F,J_1} \right) > \epsilon_0 - \epsilon_n \right) > 1 - 2e^{-\tau}$. It is easy to see now that since $\epsilon_n \to 0$ with $n \to \infty$, the gap $\tilde{\epsilon}_n = \epsilon_0 - \epsilon_n \to \epsilon_0 > 0$. So this establishes (ii).

(iii) Corollary 17 along with Assumption (A1) gives us that for any $J$,

$$P^n \left( D \in (\mathcal{X} \times \mathcal{Y})^n : |\mathcal{R}_{L,P} \left( f_{D,F,J} \right) - \mathcal{R}_{L,P,F}^{*} | < \eta_n \right) > 1 - e^{-\tau},$$

where $\eta_n = 4B\sqrt{2\tau n^{-1/2} + 20/3B\tau n^{-1} + 8K_1 a^p n^{-1/2}}$ and hence the ‘if’ condition of (iii) follows.
Now for $J_1 \in \tilde{J}$ and $J_2 \notin \tilde{J}$ we have $R^*_{L,P,F,J_2} - \epsilon_0 \geq R^*_{L,P,F,J_1}$, and the ‘only if’ condition of (iii) follows since

$$P^n (D \in (X \times Y)^n : R_{L,P} (f_{D,F,J_2}) > R^*_{L,P,F,J_1} + \epsilon_0 - \eta_n) > 1 - e^{-\tau},$$

since $\eta_n \to 0$ with $n \to \infty$, so the gap $\tilde{\epsilon}_n = \epsilon_0 - \eta_n \to \epsilon_0 > 0$.

SUPPLEMENTARY MATERIAL

Supplement A: Matlab Code

(http://www.bios.unc.edu/ kosorok/RFE.html). Details on the codes are given in the html page.

REFERENCES


