The Annals of Statistics
2019, Vol. 47, No. 1, 497–526
https://doi.org/10.1214/18-AOS1696

FEATURE ELIMINATION IN KERNEL MACHINES IN MODERATELY HIGH DIMENSIONS[1]

BY SAYAN DASGUPTA*, YAIR GOLDBERG† AND MICHAEL R. KOSOROK*

University of North Carolina at Chapel Hill and University of Haifa†*

We develop an approach for feature elimination in statistical learning with kernel machines, based on recursive elimination of features. We present theoretical properties of this method and show that it is uniformly consistent in finding the correct feature space under certain generalized assumptions. We present a few case studies to show that the assumptions are met in most practical situations and present simulation results to demonstrate performance of the proposed approach.

1. Introduction.

With recent advancement in data collection and storage, we have large amounts of information at our disposal, especially with respect to the number of explanatory variables or "features." When these features contain redundant or noisy information, estimating the functional connection between the response and these features can become quite challenging, and that often hampers the quality of learning. One way to overcome this is by finding a smaller set of features or explanatory variables that can perform the learning task sufficiently well.

In this paper, we discuss feature elimination in statistical learning with kernel machines. Kernel machines (KM), which we review in Section 2, are a collection of optimization algorithms for learning in pattern analysis and regression. These algorithms attempt to minimize a regularized version of the empirical risk over a reproducing kernel Hilbert space (RKHS) of functions defined on the input space $\mathcal{X}$ [referred to as $H(\mathcal{X})$] for a given loss function L. Of the many KM algorithms, the linear support vector machine (SVM), where each $f \in H(\mathcal{X})$ is a linear combination of the attributes in $\mathcal{X}$, is the simplest case. In general, the term kernel machine is reserved for the more general version of the problem with nonlinear transformation of the feature space. The popularity of these algorithms is motivated by the fact that these are easy-to-compute techniques that enable estimation under weak or no assumptions on the distribution [see Steinwart and Christmann (2008)]. The standard KM decision function typically utilizes all the input features.

Received December 2015; revised November 2017.

[1] Supported in part by Israel Science Foundation Grant 1308/12 and supported in part by United States NCI Grant PO1 CA142538.

MSC2010 subject classifications. 62G20.

Key words and phrases. Kernel machines, support vector machines, variable selection, recursive feature elimination.

However, the prediction quality of these methods often suffers under high noise-to-signal ratio, even if the dimension of the input space is only moderately high. In Section S5 of the Supplementary Material [Dasgupta, Goldberg and Kosorok (2018)], we present an example (see Table S5.1) for a nonlinear classification with only ten features, of which only two are relevant. We see that applying a meaningful feature selection method there can cut classification error in half, from 31% to about 12–14%, for a sample size of $n = 100$. It is thus a very important task to be able to select the correct feasible set of input features on which the learning can be applied.

Multiple methods have been proposed for the case when the assumed functional form of the decision rule is linear. For example, many *embedded methods*[2] with different modifications have been proposed; such as redefining the linear KM training to include sparsity in Weston et al. (2003), using the l_1 penalty as in Bradley and Mangasarian (1998), Zhu et al. (2003), the SCAD penalty in Zhang et al. (2006a), the l_q penalty [Liu et al. (2007)] or the elastic net [Wang, Zhu and Zou (2006)]. Although these methods have strong theoretical guarantees, they are relevant only in linear KMs (or SVMs), and become ineffectual in the framework of RKHSs with nonlinear kernels (such as the Gaussian RBF kernel). The widely applicable linear version is the most popular and well known of the general class of KM problems, and has been the focus of most of the prevalent feature selection techniques. Nonlinear versions of the algorithm have however become increasingly important recently, and many statistical learning problems explicitly depend on functional relationships that are strictly nonlinear in nature, for example, in protein classification [see Leslie et al. (2004)], in image classification [see Chapelle, Haffner and Vapnik (1999)], etc. Thus, feature selection for nonlinear kernel machines is the key focus for us in this paper.

A few techniques do exist that can be effectively catered to the nonlinear kernel machines framework. For example, Guyon et al. (2002) developed a *wrapper*[3]-based backward elimination procedure by recursively computing the learning function, known widely as recursive feature elimination (RFE). Although RFE was developed as an off-the-shelf technique for linear KMs, the authors included an analogous formulation for the nonlinear transformed space as well. 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 function after deletion of each of the features remaining in the model, and removes the one with the lowest change in such norm, thus performing an implicit ranking of features. A number of approaches have been developed inspired by RFE [see Rakotomamonjy (2003), Tang, Zhang and Huang (2007), Mundra and Rajapakse

[2] Methods that construct the learning algorithm in a way to include feature elimination as an in-built phenomenon.

[3] Methods that use the learning method itself to score feature subsets.

(2010)]. RFE has been studied extensively in the bioinformatics and computer science literature [see, e.g., Zhang et al. (2006b), Aksu et al. (2010), Aksu (2014)]. It has also been used for feature selection in many recent applications [see, e.g., Hu et al. (2010), Hidalgo-Muñoz et al. (2013), Louw and Steel (2006)]. Recently, a new multistage *embedded* optimization method has been proposed [see Allen (2013)]. However, the key drawback of most of these methods is that their theoretical properties have never been studied rigorously.

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 modifying RFE to create a recursive technique that can be validated as a theoretically sound procedure for feature elimination in kernel machines. Our main contributions include:

(1) We *develop a theoretical framework* that can validate feature elimination in KMs. For example, since optimization is restricted within the RKHS $H(\mathcal{X})$ in KMs, one important task here is to redefine $H(\cdot)$ on any lower dimensional domain of $\mathcal{X}$, so that it retains its RKHS properties.

(2) We *modify the criterion for deletion and ranking of features* from Guyon et al.'s RFE, and call it the risk-RFE algorithm. The ranking of the features here is based on the lowest difference observed in the regularized empirical risk after removing each feature from the existing model. This is done to enable theoretical consistency.

(3) We establish *asymptotic consistency of the modified risk-RFE algorithm* in finding the "correct" feature space, both when the *dimension of the input space is fixed*, and when the *dimensionality is allowed to grow with the sample size*. We discuss at length the necessary conditions for achieving consistency under both setups, and for the latter, establish a range of allowed rates of dimensionality growth that can guarantee consistency. We believe these are some of the first theoretical results on feature selection in kernel machines.

(4) We discuss the applicability of our methods in some *important learning problems, including image classification.*

(5) We discuss a practical method of using our algorithm to *select an optimal feature set* for learning.

The paper is organized as follows: In Section 2, we present a short summary of the problem, the proposed feature elimination algorithm for kernel machines, and the main theoretical results of the article. In Section 3, we present an in depth analysis of the various assumptions for the risk-RFE algorithm and discuss their implications. In Section 4, we prove our main results under the most general setting, following which several case studies are discussed in-depth in Section 5. In Section 6, we provide simulation results to demonstrate how risk-RFE can be used in intelligent selection of features, and assess its performance in various settings of nonlinearity. In Section 7, we apply our algorithm to several applied data settings, both in classification and regression. A discussion is provided in Section 8, while additional materials and detailed proofs are given in the Supplementary Material [Dasgupta, Goldberg and Kosorok (2018)], along with a link to the software codes.

2. The risk-recursive feature elimination algorithm (risk-RFE).

In this section, we summarize the main findings of the paper. We first briefly describe the relevant problem, along with its mathematical formulation, and then follow up with the risk-RFE algorithm and our main consistency results.

2.1. *The problem description.* Given a set of training data $D = \{(X_1, Y_1), \dots, (X_n, Y_n)\} \in (\mathcal{X} \times \mathcal{Y})^n$, a typical goal in statistical learning is to estimate a rule that can be used to predict Y for a given input feature vector X. In kernel machines, this is done by minimizing a regularized version of the empirical risk (for a given loss function L) of functional rules obtained from a reproducing kernel Hilbert space (RKHS). An RKHS H is typically represented by a bi-linear function $k(\cdot, \cdot)$, and for a given transformation ϕ of the feature space, the appropriate RKHS H is the one with kernel k satisfying $k(x_1, x_2) = \langle \phi(x_1), \phi(x_2) \rangle_H$.

Linear KM for binary classification with the hinge loss $L_{\mathrm{HL}}(X, Y, f(X)) = \max(1 - Yf(X), 0)$ is the most popular version of the kernel machines algorithm, but for the untransformed feature space. It has been extensively studied, under several feature selection methods (especially, the L_p penalized forms of the algorithm as they are easily interpretable in the linear case). However, feature selection in general kernel machines is still a relatively new area of research, and one key goal here is to lay the theoretical foundation of a feature selection method in nonlinear KMs.

2.2. *Mathematical formulation.* The notation and the oracle bounds used throughout the paper will closely follow Steinwart and Christmann (2008) (hereafter abbreviated SC08). Consider the measurable space $(\mathcal{X}, \mathcal{A}, P^d_{\mathcal{X}})$ such that $\mathcal{X} \subseteq B \subset \mathbb{R}^d$ is a valid metric space, with B a d-dimensional open Euclidean ball centered at 0. Let $\mathcal{Y}$ be a closed subset of $\mathbb{R}$ and $P^d_{\mathcal{X}\times\mathcal{Y}} := P^d$ be a measure on $\mathcal{X} \times \mathcal{Y}$, such that $P^d_{\mathcal{X}}$ is a restriction of P^d on $\mathcal{X}$, and let d_0 denote the number of relevant features in $\mathcal{X}$. We start by defining the kernel machine algorithm in its most general forms.

REMARK 1. The L-risk (for a given loss function L) of the measurable function f is given as $\mathcal{R}_{L,P}(f) = E_P[L(X, Y, f(X))]$. The Bayes' risk $\mathcal{R}^*_{L,P}$ is defined as $\inf_f \mathcal{R}_{L,P}(f)$, where the infimum is taken over $\mathcal{L}_0(\mathcal{X}) = \{f : \mathcal{X} \mapsto \mathbb{R}, f \text{ is measurable}\}$, the set of all measurable functions. A function f^*_P that achieves this infimum is called a Bayes' decision function. Let $\mathcal{F}$ be a given optimization space, and $f_{P,\mathcal{F}} = \arg\min_{f\in\mathcal{F}} E_P[L(X, Y, f(X))] = \arg\min_{f\in\mathcal{F}} \mathcal{R}_{L,P}(f)$ be the minimizer of infinite-sample risk within $\mathcal{F}$. We denote this minimal risk as $\mathcal{R}^*_{L,P,\mathcal{F}} = \mathcal{R}_{L,P}(f_{P,\mathcal{F}})$.

REMARK 2. The loss L is called convex when $L(x, y, \cdot)$ is convex for every $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. It is also locally Lipschitz continuous if for every $a > 0$, $\sup_{x\in\mathcal{X},y\in\mathcal{Y}} |L(x, y, s) - L(x, y, \acute{s})| < c_L(a)|s - \acute{s}|, s, \acute{s} \in [-a, a]$ for a given local constant $c_L(\cdot)$. Note that the results developed here are equally valid for regression under certain regular assumptions on $\mathcal{Y}$.

KERNEL MACHINE (KM): Consider now a loss function L, which is convex, locally Lipschitz continuous and measurable, and H (note that H is a special form of the optimization space $\mathcal{F}$), a separable RKHS of a measurable kernel k on $\mathcal{X}$, and fix a $\lambda > 0$. The *general KM solution* is the function $f_{P,\lambda,H} \in H$ that satisfies

$$f_{P,\lambda,H} = \arg\min_{f\in H} \lambda\|f\|_H^2 + \mathcal{R}_{L,P}(f). \tag{1}$$

For the observed data D, the *empirical KM decision function* $f_{D,\lambda,H}$ is then given as

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

REMARK 3. The kernel k of the RKHS H is a unique, real-valued symmetric function $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$. The kernel k has the reproducing property that $f(x) = \langle f, k(\cdot, x)\rangle_H$ for all $f \in H$, and all $x \in \mathcal{X}$, where $\langle\cdot,\cdot\rangle_H$ is the inner product induced by H. Moreover, we also have $k(\cdot, x) \in H$, for all $x \in \mathcal{X}$.

2.3. *The feature elimination algorithm.* Limitations of Guyon et al.'s RFE as a margin-maximizing feature elimination were studied explicitly in Aksu et al. (2010). Hence, as opposed to Guyon et al., who used the Hilbert space norm $\lambda\|f\|_H^2$ to eliminate features recursively, we use the entire objective function (the regularized empirical risk) for deletion.

For a probability measure Q and the optimization space $\mathcal{F}$, define the regularized Q-risk as

$$\mathcal{R}^{\text{reg},\lambda}_{L,Q,\mathcal{F}}(f) = \lambda\|f\|_{\mathcal{F}}^2 + \mathcal{R}_{L,Q}(f). \tag{3}$$

Also define the restricted space $\mathcal{F}^J$ (often referred to as a pseudo-subspace of $\mathcal{F}$) as follows.

DEFINITION 1. Let J be a set of indices $J \subseteq \{1, 2, \dots, d\}$. Then for a given functional space $\mathcal{F}$, define $\mathcal{F}^J = \{g : g = f \circ \pi^{J^c}, \forall f \in \mathcal{F}\}$, where π^{J^c} is the projection map that takes element $x \in \mathbb{R}^d$ and maps it to $x^J \in \mathbb{R}^d$, by substituting elements in x indexed in the set J, by zero, and leaving the remaining elements unchanged.

REMARK 4. Note that we can subsequently define the space $\mathcal{X}^J = \{\pi^{J^c}(x) : x \in \mathcal{X}\}$. Thus the above formulation allows us to create lower dimensional versions of a given functional space $\mathcal{F}$.

REMARK 5. One important aspect of the problem is the dimensionality (dimension d of $\mathcal{X}$ and the number of signals d_0). With a recent surge in interest in the high dimensional version of standard problems, wherein the asymptotic properties of the design size d are studied along with those of the sample size n, it is important that we evaluate risk-RFE in the same light as well. In this article, we consider both the standard fixed dimensional setting, as well as the setting where d and d_0 grow with n, but with the restriction $d > d_0 > 0$. The varying dimensional setting requires additional technical details beyond the standard one, and hence in the following sections we will study our algorithm in the fixed dimensional setting first (often P will replace P^d in such cases), followed by modifications and other technical requirements for the varying dimensional setting.

For most practical purposes, the algorithm will be used under the paradigm of fixed dimension only, as shown below. For now, assume d and d_0 (the number of relevant features) to be fixed constants. The risk-RFE algorithm, defined for the parameters $\{\lambda_n, \delta_n\}$ is given as the following.

ALGORITHM 1 (risk-RFE in the fixed dimensional setting). For a given RKHS H, we start off with $J \equiv [\cdot]$ empty and let $Z \equiv [1, 2, \dots, d]$.

STEP 1: In the kth iteration, choose feature $i_k \in Z \setminus J$ which minimize

$$\mathcal{R}^{\mathrm{reg},\lambda_n}_{L,D,H^{J\cup\{i\}}}(f_{D,\lambda_n,H^{J\cup\{i\}}}) - \mathcal{R}^{\mathrm{reg},\lambda_n}_{L,D,H^{J}}(f_{D,\lambda_n,H^{J}}), \tag{4}$$

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

Continue this until the difference

$$\min_{i \in Z\setminus J} \mathcal{R}^{\mathrm{reg},\lambda_n}_{L,D,H^{J\cup\{i\}}}(f_{D,\lambda_n,H^{J\cup\{i\}}}) - \mathcal{R}^{\mathrm{reg},\lambda_n}_{L,D,H^{J}}(f_{D,\lambda_n,H^{J}})$$

becomes larger than a pre-determined quantity δ_n, and output J as the set of indices for the features to be removed from the model.

The quantity λ_n is the tuning parameter associated with the Hilbert norm of the functional rule and controls over-fitting. The quantity δ_n is the other tuning parameter controlling feature selection. It is given as the maximum limit of increment allowed in the objective function during two successive steps of feature selection in the risk-RFE algorithm. The oracle choices for the parameters λ_n and δ_n will be discussed in Section 2.6. Now, consider the situation when both d, the dimension of $\mathcal{X}$, and (potentially) d_0, the number of relevant features, go to infinity with n. The modified feature selection algorithm for the varying dimensional one is given as the following.

ALGORITHM 2. Replace the stopping condition in Algorithm 1 from δ_n to $\delta_n^{P^d}(d - |J|)$, where $\delta_n^{P^d}(\cdot)$ is a positive monotone decreasing function.

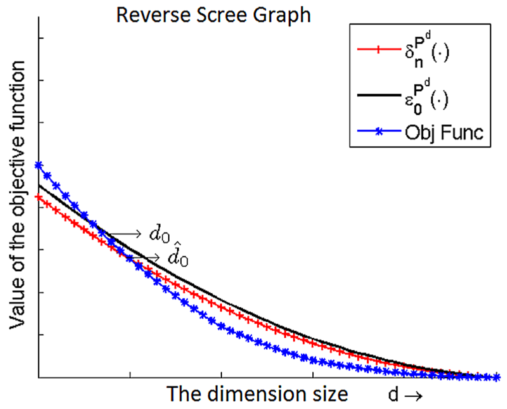

FIG. 1. *Stopping rule for the modified algorithm in the growing design size setting*: *A potential case.*

The only modification of the algorithm lies in the stopping rule. The fixed tuning parameter δ_n in the fixed design problem is replaced by the function $\delta_n^{P^d} : \{1, \dots, d\} \mapsto \mathbb{R}$. Figure 1 shows a visual representation of the stopping condition in this case. Extending the rationale from the fixed design case, $\delta_n^{P^d}$ acts as a functional upper bound for the difference of the regularized risks (the objective function for risk-RFE). In other words, at iteration i of the algorithm, $\delta_n^{P^d}(d-i)$ acts as the maximal allowance that the difference of regularized risks (between models at subsequent iterations) can attain at this iteration. Hence, the point where this difference finally jumps above $\delta_n^{P^d}(\cdot)$, is where our algorithm is stopped, and the features left in the model are retained as potential signals.

One thing to note is that Algorithm 2 is a natural extension of Algorithm 1, presented here to illustrate the changes of the underlying dynamics in the varying dimensional setting. For most practical purposes, dimension d of the covariate space will be fixed, and we will really use Algorithm 1, presented earlier. In that case, $\delta_n^{P^d}(\cdot)$ reduces to the fixed δ_n, the optimal choice of which, in practical settings, will be discussed in Section 6.1.

2.4. *Number of features removed in each iteration of risk-RFE.* In Algorithm 1 above, we discussed removing only one feature at each iteration. Similarly, one can also consider removing multiple features (say k) in a single iteration. In that case, the k indices that produce the k smallest values of the objective function (4) are removed in that iteration. This number can also be defined adaptively, with different numbers of features removed in different iterations of the algorithm. For simplicity, we have set it to 1 for our theory, but in numerical simulations we often define it adaptively to speed up computations.

2.5. *Heuristics on why risk-RFE is consistent.* Consider risk minimization in the fixed dimensional case given the optimization space $\mathcal{F}$, and let our goal be to find a solution $f \in \mathcal{F}$ that minimizes a given empirical criterion (such as "regularized empirical" risk in kernel machines). If it so happens that the minimizer of the infinite-sample risk resides in a space spanned by a lower dimensional subspace of $\mathcal{X}$ (say $\mathcal{X}^*$), then it may actually suffice to find the empirical minimizer over the restriction of $\mathcal{F}$ on $\mathcal{X}^*$. To avoid over-fitting, this indeed becomes necessary. Hence the motivation for our algorithm stems from the following two heuristics: (a) if any feature is superfluous for the given problem, then given all other features in the model, its contribution to the functional relationship between the output variable and the feature space should only be due to random fluctuations, and should therefore be small. Thus the incremental risk associated with a solution in the subspace defined by ignoring this surplus feature, when compared to the solution in the original feature space, should be minimal; (b) if a signal is removed from the model, we will expect this incremental risk to be substantial, and greater than some unknown oracle specific to the design of the problem. We will refer to this quantity as ε_0.

2.6. *Consistency results for risk-RFE.* The main results of this paper will be summarized as consistency statements for our algorithm under two separate paradigms. These will be defined by a set of regularity conditions and different underlying assumptions, which we summarize below along with the main results in the form of three separate theorems.

Let us first start with the *consistency statements* that we want to establish for our algorithm. For the risk-RFE Algorithm 1 for kernel machines with a RKHS H and tuning parameters (δ_n, λ_n), we want to show that the following statements hold:

(CS1) The risk-RFE algorithm finds the correct lower dimensional subspace of the input space $\mathcal{X}$ with probability tending to 1.

(CS2) The function chosen by risk-RFE achieves the minimal risk within the given RKHS H asymptotically.

REMARK 6. We will denote this correct lower dimensional subspace by $\mathcal{X}^{J_*}$, such that variables that carry no signal are indexed by the set J_*, which means $\mathcal{X}^{J_*}$ contains only the signals (or relevant variables). This will be formally established by equation (8) in Section 3.1.

Next, we provide a set of regularity conditions that will be necessary for the consistency arguments to hold. Before that, let us recall that for a given metric space (T, d) and for any integer $n \geq 1$, the ith entropy number of (T, d) is defined as

$$e_i(T, d) := \inf\left\{\varepsilon > 0 : \exists s_1, \dots, s_{2^{i-1}} \in T \text{ such that } T \subset \bigcup_{j=1}^{2^{i-1}} B_d(s_j, \varepsilon)\right\}, \tag{5}$$

where $B_d(s,\varepsilon)$ is the closed ball of radius ε centered at s, with respect to the metric d. If $S : E \mapsto F$ is a bounded linear operator between normed spaces E and F, we write $e_i(S) = e_i(SB_E, \|\cdot\|_F)$, where B_E is the unit closed ball in E.

The *regularity conditions* are:

(RC0) Let H be as defined before, and assume L satisfies $L(x,y,0) \leq 1$ for all $(x,y) \in \mathcal{X} \times \mathcal{Y}$, and the kernel k on $\mathcal{X}$ is such that $\|k\|_\infty \leq 1$. Also let $\{\lambda_n\} \in [0,1]$ be such that $\lambda_n \to 0$ and $\lim_{n\to\infty} \lambda_n n = \infty$.

REMARK 7. Conditions $L(x,y,0) \leq 1$, and $\|k\|_\infty \leq 1$ above are assumed for simplicity and equivalent conditions such as $L(x,y,0) \leq M$ and $\|k\|_\infty \leq k_{\text{sup}}$ for constants $M, k_{\text{sup}} > 1$ can also guarantee the desired result.

Regularity conditions (RC0) are standard conditions that are assumed for deriving oracle bounds for KM solutions [e.g., see Steinwart and Christmann (2008)]. Now note that the approximation error $A_2^{J_*}(\lambda)$ is defined as the difference between the regularized risk of the oracle minimizer of such and the minimum risk achieved within the space H^{J_*} (see Glossary for a precise definition). Now we define the next set of regularity conditions (RC1).

(RC1) Assume, for fixed $n \geq 1$, there exist constants $a \geq 1$ and $p \in (0,1)$ such that the following *entropy condition* holds:

$$\mathbb{E}_{D_\mathcal{X} \sim P_\mathcal{X}^n} e_i\big(id : H \mapsto L_\infty(D_\mathcal{X})\big) \leq a i^{-\frac{1}{2p}}, \qquad i \geq 1, \tag{6}$$

where $\mathbb{E}_{D_\mathcal{X} \sim P_\mathcal{X}^n}$ is the expectation with respect to the product measure $P_\mathcal{X}^n$ for data $D_\mathcal{X} \equiv \{\mathcal{X}_1, \ldots, \mathcal{X}_n\}$ being i.i.d. copies of $\mathcal{X} \sim P_\mathcal{X}$. We also assume that there exists $c > 0$ and $\beta \in (0,1]$ such that $A_2^{J_*}(\lambda) \leq c\lambda^\beta$ for all $\lambda \geq 0$.

Regularity conditions (RC1) are necessary for establishing consistency when d and d_0 are fixed. These are additional conditions on the entropy of the RKHS H and the approximation error that were also naturally assumed in Steinwart and Christmann (2008) to derive the oracle bounds under the fixed dimensional setting.

REMARK 8. Sometimes, it will be useful to replace the entropy condition (6) by a slightly different condition (7) given below [as will be seen in the Supplementary Material (Section S8.3)], under which, all our statements will continue to hold. For fixed $n \geq 1$, we assume there exists constants $a \geq 1$ and $p \in (0,1)$ such that for any $J \subseteq \{1,2,\ldots,d\}$,

$$\mathbb{E}_{D_\mathcal{X} \sim P_\mathcal{X}^n} e_i\big(id : H^J \mapsto L_2(D_\mathcal{X})\big) \leq a i^{-\frac{1}{2p}}, \qquad i \geq 1. \tag{7}$$

Now we will briefly introduce the last set of regularity conditions.

(RC2) When both d and d_0 grow with sample size n, different bounds on the entropy and the approximation error are needed than those given in (RC1), and we will call them regularity conditions (RC2). Because they involve further technical details, we postpone their actual definition to Section 3.5.

Let us now note the different *conditions* under which consistency for the risk-RFE algorithm can be attained.

CONDITION 1 (Existence of a null model). Heuristically, this condition states that there exists a nontrivial subset of the covariate set, which is the correct set of features for the given problem. The precise definition is given by equation (8) in Section 3.1.

CONDITION 2A (Nestedness/denseness). This condition assumes that the reproducing kernel Hilbert space H admits the nestedness or denseness property. H admits the nestedness property if $H^{J_2} \subseteq H^{J_1}$ holds whenever the index sets J_1, J_2 are such that $J_1 \subseteq J_2$. H admits the denseness property when it is dense in a functional class that admits the nestedness property.

We will define these properties more explicitly in Section 3.2. Most common optimization spaces satisfy either nestedness (e.g., linear RKHS produced by the Euclidean inner product) or denseness (e.g., RKHS generated by the Gaussian RBF kernel), and will be sufficient for consistency statements (CS1) and (CS2) to hold. We are now ready to give our first result.

THEOREM 1. *Assume that d, the dimension of $\mathcal{X}$, and the number of relevant features d_0, are fixed. Then for $\delta_n = \varepsilon_0 - \mathcal{R}(n)$, where $\mathcal{R}(n) > 0$ is such that $\mathcal{R}(n)n^{\frac{\beta}{2\beta+1}} \to C$ for some $C \in (0, \infty)$ and β is the constant appearing in* (RC1), *consistency statements* (CS1) *and* (CS2) *hold under Conditions* 1 *and* 2a, *and regularity conditions* (RC0) *and* (RC1).

REMARK 9. We heuristically defined ε_0 in Section 2.5, but we give a more formal definition in Section 3.4 when we discuss the finite gap condition in more detail.

REMARK 10. In Section 3.3.3, we will introduce a more general version of Theorem 1 referred to as Theorem 1A, and finally in Section 4.1, we will provide a proof for Theorem 1A.

The above theorem assumes that d, the dimension of $\mathcal{X}$, and d_0, the number of relevant features, are fixed. In this scenario, we can always define a "gap" as the infimum difference of minimized risks attained (for the problem) in the "correct" restriction of H and any of its incorrect restrictions. This will be further elaborated in Remark 13, and in the definition of Assumption (A2) in Section 3.4, which becomes a natural extension of the nestedness/denseness condition in this setting. But when both d and d_0 grow with sample size n, the idea of a fixed gap is no longer a natural consequence, because it might diminish and shrink to 0 as d tends to infinity. Hence, for our consistency results to hold in that setting, we need an additional condition given below:

CONDITION 2B (Asymptotically vanishing gap). This condition says that when the design size d grows with sample size n, the infimum of the difference between minimized risk attained within the "correct" restriction of H and any of its incorrect restrictions, shrinks asymptotically to 0.

The precise definition of the asymptotically vanishing gap condition (Condition 2b) will be given below as (A2*) in Section 3.5. Before we give our third and final result of this paper, note that as $\delta_n^{P^d}(\cdot)$ replaces δ_n in Algorithm 2 for the varying dimensional setting, the quantity ε_0 is also replaced by the functional $\varepsilon_0^{P^d}(\cdot)$ by similar logic. We will study these quantities in more detail in Sections 3.2 and 3.3.

THEOREM 2. *Assume that d, the dimension of $\mathcal{X}$, and d_0, the number of relevant features, grow with n. Then for some $\gamma \in (0, \frac{\beta}{2\beta+1}]$ where β is the constant appearing in* (RC1), *and $\delta_n^{P^d}(\cdot) = \varepsilon_0^{P^d} - \mathcal{P}(n)$ such that $\mathcal{P}(n)n^\gamma \to K$ for some $K \in [0, \infty)$, consistency statements* (CS1) *and* (CS2) *hold under Conditions* 1, 2a *and* 2b, *and regularity conditions* (RC0) *and* (RC2).

REMARK 11. We will introduce a more general version of Theorem 2 (referred to as Theorem 2A) in Section 3.3.3. The proof of Theorem 2A will be provided as modifications to the proof of Theorem 1A in Section 4.2.

3. Assumptions for risk-RFE.

In this section, we discuss the above conditions in further details. We will verify some of them for a few examples later in Section 5. We start by looking at the "existence of a null model" condition.

3.1. *Existence of a null model.*

Let $\mathcal{F}$ be a general optimization space. Existence of a null model means that there exists an index set J_* such that

$$\mathcal{R}^*_{L,P,\mathcal{F}} = \mathcal{R}^*_{L,P,\mathcal{F}^{J_*}}. \tag{8}$$

REMARK 12. Note that the above does not claim the uniqueness of J_*, but that for any set of covariates with the above property (8), there always exists a (possibly nonunique) maximal set satisfying it. Also observe that J_* can be empty when $f_{P,\mathcal{F}}$ spans through all the sub-dimensions of $\mathcal{F}$, and that would mean that we need the entire space $\mathcal{F}$ for the optimization.

3.2. *The nestedness/denseness property.* Here, we discuss the scope of risk-RFE when the functional space admits certain nice properties such as nestedness or denseness.

3.2.1. *Nested spaces in risk minimization.* In risk minimization, we say $\mathcal{F}$ admits the nestedness property, if for a pair of index sets $J_1, J_2 \in \{1, 2, \dots, d\}$ with $J_1 \subseteq J_2$, we have $\mathcal{F}^{J_2} \subseteq \mathcal{F}^{J_1}$. This translates to admitting nested inequalities of the form $\mathcal{R}^*_{L,P,\mathcal{F}^{J_1}} \leq \mathcal{R}^*_{L,P,\mathcal{F}^{J_2}}$. One simple example is the linear space, $\mathcal{F} = \{f(x_1, \dots, x_d) = \sum_i a_i x_i : |a_i| \leq M, M < \infty\}$.

In general, RKHSs need not be nested within each other. We will see below, however, that dot-product kernels do have this property.

LEMMA 3. *Dot product kernels produce nested RKHSs.*

The proof is given in the Supplementary Material (Section S8.1). Dot product kernels (e.g., linear kernels) appear quite regularly in KM problems. Other kernels may also display the nestedness property.

3.2.2. *Dense spaces in risk minimization.* We say $\mathcal{F}$ admits the denseness property, if it is dense in a functional class that admits the nestedness property [e.g., the space of bounded measurable functions $\mathcal{L}_\infty(\mathcal{X})$ or the space of continuous and bounded functions $\mathcal{C}(\mathcal{X})$]. Note that all universal kernels produce RKHSs that are dense in $\mathcal{C}(\mathcal{X})$ and attain the Bayes' risk (i.e., $\mathcal{R}^*_{L,P,\mathcal{F}} = \mathcal{R}^*_{L,P}$) if the loss function is convex and locally Lipschitz continuous, and $\mathcal{X}$ is compact. All nontrivial radial kernels (e.g., Gaussian RBF kernel) share this property as well [see Micchelli, Xu and Zhang (2006)], and hence this is a fairly typical framework for KM problems.

3.2.3. *Implications of the nestedness/denseness condition.* In spaces which admit the nestedness property, the existence of a null model condition is equivalent to stating that there exists a minimizer of infinite-sample risk in $\mathcal{F}$, which also belongs to $\mathcal{F}^{J_*}$. This then trivially implies $\mathcal{R}^*_{L,P,\mathcal{F}^J} = \mathcal{R}^*_{L,P,\mathcal{F}^{J_*}}$ whenever $J \subseteq J_*$. If $\mathcal{F}$ is now dense in a functional space $\mathcal{G}$ admitting the nestedness property [e.g., $\mathcal{L}_\infty(\mathcal{X})$], then by Lemma S3 of the Supplementary Material (Section S1.1), $\mathcal{F}^J$ is dense in $\mathcal{G}^J$ for any $J \in \{1, 2, \dots, d\}$. Hence "denseness" does not necessarily imply "nestedness," but we do have the "almost nestedness" property in the sense that any function $g \in \mathcal{F}^{J_2}$ can be well approximated by a sequence of functions $\{f_n\} \in \mathcal{F}^{J_1}$ for $J_1 \subseteq J_2$. This actually implies (8) (given above) for any $J \subseteq J_*$.

REMARK 13. Note that when d and d_0 are fixed, we can define $\varepsilon_0 = \min_{J_\circ \not\subseteq J_*} \mathcal{R}^*_{L,P,\mathcal{F}^{J_\circ}} - \mathcal{R}^*_{L,P,\mathcal{F}^{J_*}}$. The above then means that $\mathcal{R}^*_{L,P,\mathcal{F}^{J_\circ}} \geq \mathcal{R}^*_{L,P,\mathcal{F}^{J_*}} + \varepsilon_0$ holds whenever $J_\circ \not\subseteq J_*$, and J_* is unique.

REMARK 14. The "nestedness structure" is essentially different from the nested model setup in Tsybakov (2004). Tsybakov (2004) started with a predecided nested sequence of classifier sets (or models) and obtained a solution from each of these classifier sets. In contrast, here we have a graph of nested models that can include many subtrees in the sense that in every intermediate step, we are presented with multiple models within the parent model. We select the best classifier from each of these models and opt for the one among them obtaining the best performance.

3.3. *The risk equivalence condition.* In this section, we propose an alternative and a more general condition, called "the risk equivalence condition," that can replace Condition 2a in Theorem 1. This allows us to move beyond the premise that $\mathcal{F}$ admits the nestedness or denseness property, and allows risk-RFE to perform consistently under more general setups. Note that the risk equivalence condition is indeed necessary for establishing theoretical consistency for a backward selection algorithm in the spirit of risk-RFE.

CONDITION 2A* (Risk equivalence). This condition heuristically says that there exists at least one sequence of subspaces, starting from $\mathcal{X}$ down to the correct feature space $\mathcal{X}^{J_*}$, such that the minimum infinite-sampled (or oracle) risk attained by the restriction of $\mathcal{F}$ on each of these subspaces is the same. That is, $\mathcal{F}$ satisfies property (A1) given below:

(A1) Let J_* be as defined in (8). Then for any pair (d_1, d_2) satisfying $d_1 \le d_2 \le d - d_0$, there exist J_{d_1} and J_{d_2} with $J_{d_1} \subseteq J_{d_2} \subseteq J_*$ and $|J_{d_1}| = d_1$ and $|J_{d_2}| = d_2$, such that $\mathcal{R}^*_{L,P,\mathcal{F}^{J_*}} = \mathcal{R}^*_{L,P,\mathcal{F}^{J_{d_1}}} = \mathcal{R}^*_{L,P,\mathcal{F}^{J_{d_2}}}$.

REMARK 15. This condition is a weaker version of the nestedness/denseness condition, which means that whenever $\mathcal{F}$ satisfies nestedness or denseness, it automatically satisfies the risk equivalency property (A1), as can be seen from our discussions in Section 3.2.3.

REMARK 16. Assumption (A1) says that there exists a "path" from the original input space $\mathcal{X}$ to the correct lower dimensional space $\mathcal{X}^{J_*}$ such that each of the spaces $\mathcal{F}^J$s along this "path" obtains the same minimized risk. Note that this path need not be unique, and there can be more than one path going down to a given J_*.

The following examples show that assumption (A1) is in fact necessary for a well-defined backward selection algorithm to work in the absence of nestedness or denseness.

3.3.1. *Necessity for existence of a path in* (A1).

EXAMPLE 1. Consider the following 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 $\mathcal{F}$ be given as $\{c_1X_1 + c_2X_2, c_1, c_2 > 1\}$, and note that $\mathcal{F}$ neither admits the nestedness property nor is dense in $\mathcal{L}_\infty(\mathcal{X})$ or in any other space admitting the nestedness property. Let us consider the squared error loss, that is, $L(x, y, f(x)) = (y - f(x))^2$. By Definition 1 in Section 2.3, $\mathcal{F}^{\{1\}} = \{c_2X_2, c_2 > 0\}$ and $\mathcal{F}^{\{2\}} = \{c_1X_1, c_1 > 0\}$ 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 $\mathcal{R}_{L,P}(f_{P,\mathcal{F}^{\{2\}}})$ are strictly positive. Thus 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). This shows that although the correct lower dimensional space may have minimized risk the same as that of the original, if there exists no path going down to that space, a backward selection algorithm will never find it.

3.3.2. *Necessity for equality in* (A1). The following example shows that equality in (A1) cannot be replaced by "$\leq$."

EXAMPLE 2. Consider another empirical risk minimization framework and assume (A1) holds with equality replaced by "$\leq$." Let $Y \sim U(-1, 1)$ and $X \subset \mathbb{R}^3$ such that $Y = X_3 = X_2 + 1 = X_1 - 1$. Let $\mathcal{F} = \{c_1X_1 + c_2X_2 + c_3X_3, c_1, c_2, c_3 \geq 1\}$, and let the loss function be squared error loss. Now by definition, $\mathcal{F}^{\{1\}} = \{c_2X_2 + c_3X_3, c_2, c_3 \geq 1\}$, $\mathcal{F}^{\{2\}} = \{c_1X_1 + c_3X_3, c_1, c_3 \geq 1\}$, $\mathcal{F}^{\{3\}} = \{c_2X_2 + c_1X_1, c_1, c_2 \geq 1\}$, $\mathcal{F}^{\{1,2\}} = \{c_3X_3, c_3 \geq 1\}$, $\mathcal{F}^{\{1,3\}} = \{c_2X_2, c_2 \geq 1\}$, $\mathcal{F}^{\{2,3\}} = \{c_1X_1, c_1 \geq 1\}$, and $\mathcal{F}^{\{1,2,3\}} = \{0\}$. By simple calculations, we see that $\mathcal{R}^*_{L,P,\mathcal{F}} = \mathcal{R}^*_{L,P,\mathcal{F}^{\{1\}}} = \mathcal{R}^*_{L,P,\mathcal{F}^{\{2\}}} = 4/3$, $\mathcal{R}^*_{L,P,\mathcal{F}^{\{3\}}} = \mathcal{R}^*_{L,P,\mathcal{F}^{\{1,2,3\}}} = 1/3$, $\mathcal{R}^*_{L,P,\mathcal{F}^{\{1,3\}}} = \mathcal{R}^*_{L,P,\mathcal{F}^{\{2,3\}}} = 1$ and $\mathcal{R}^*_{L,P,\mathcal{F}^{\{1,2\}}} = 0$. Note that the correct dimensional subspace of the input space is $X^{\{1,2\}}$ and there exists paths leading to this space via $X \to X^{\{1\}} \to X^{\{1,2\}}$ (since $\mathcal{R}^*_{L,P,\mathcal{F}} = \mathcal{R}^*_{L,P,\mathcal{F}^{\{1\}}} > \mathcal{R}^*_{L,P,\mathcal{F}^{\{1,2\}}}$) or via $X \to X^{\{2\}} \to X^{\{1,2\}}$ (since $\mathcal{R}^*_{L,P,\mathcal{F}} = \mathcal{R}^*_{L,P,\mathcal{F}^{\{2\}}} > \mathcal{R}^*_{L,P,\mathcal{F}^{\{1,2\}}}$). However, there also exists a blind path $X \to X^{\{3\}}$ (since $\mathcal{R}^*_{L,P,\mathcal{F}} > \mathcal{R}^*_{L,P,\mathcal{F}^{\{3\}}}$), which does not lead to the correct subspace. So a recursive search with modified (A1) is not guaranteed to lead to the correct subspace.

3.3.3. *More general versions of Theorems* 1 *and* 2. It is worthwhile to note that our consistency statements from Theorems 1 and 2 continue to hold if the nestedness/denseness condition (Condition 2a) is replaced by the risk equivalence Condition 2a* as summarized by the following results.

THEOREM 1A. *Consistency statements* (CS1) *and* (CS2) *continue to hold under the premise of Theorem* 1, *if we replace Condition* 2a *by* 2a*.

THEOREM 2A. *Consistency statements* (CS1) *and* (CS2) *continue to hold under the premise of Theorem* 2, *if we replace Condition* 2a *by* 2a*.

REMARK 17. Note that to prove the main results under Condition 2a (Theorems 1 and 2), it is enough to have $A_2^J(\lambda) \le c\lambda^\beta$ hold for $J = J_*$ as in (RC1), but to do the same under Condition 2a* (Theorems 1A and 2A), we actually need the bound to hold for any $J \subseteq J_*$.

3.4. *The finite gap condition in the fixed design setting.* As discussed in Section 2.5, a feature is defined as a *signal* if and only if the risk of the model is inflated in its absence. Equivalently, if a feature does not contribute to the model at all, the increase in risk (regularized or nonregularized) on its removal should be inconsequential. We now formally define the quantity ε_0, which we hypothesized as a lower bound for the increment of risk when a signal is removed from the model. When the design size d is finite, note that assumption (A1) (similar to what we deduced in Remark 13) implies:

(A2) Let $\mathcal{J}_1, \mathcal{J}_2, \dots, \mathcal{J}_N$ be the exhaustive list of such paths from $\mathcal{X}$ to $\mathcal{X}^{J_*}$, and let $\widetilde{\mathcal{J}} := \bigcup_{i=1}^N \mathcal{J}_i$. There exists $\varepsilon_0 > 0$ such that whenever $J \notin \widetilde{\mathcal{J}}$, $\mathcal{R}^*_{L,P,\mathcal{F}^J} \ge \mathcal{R}^*_{L,P,\mathcal{F}^{J_*}} + \varepsilon_0$.

Equality in (A1) guarantees that the recursive search will never select an important dimension $j \in J_*^c$ for redundancy because then (A2) would be violated. Hence (A1) ensures that we follow the correct path recursively, and (A2) gives us a stopping condition to halt at the correct input space $\mathcal{X}^{J_*}$.

3.5. *The asymptotically vanishing gap condition in the varying design setting.* When d and d_0 grow with n, (A2) does not follow naturally from nestedness/denseness or risk equivalence as we saw above. A fixed gap makes sense when d is fixed, but in a varying design problem, this gap might diminish and shrink to 0 as d tends to infinity. Thus we propose the existence of a function $\varepsilon_0^{P^d}(\cdot)$, a strictly positive and monotonically decreasing function from $\{1, \dots, d\} \mapsto \mathbb{R}$, such that $\varepsilon_0^{P^d}(d - d_0)$ goes to zero in limit, when $d \to \infty$, and we modify (A2) to (A2*) below to accommodate these changing dynamics.

(A2*) Let $\mathcal{J}_1, \mathcal{J}_2, \dots$ be an enumeration of paths from $\mathcal{X}$ to $\mathcal{X}^{J_*}$, and let $\widetilde{\mathcal{J}} := \bigcup_i \mathcal{J}_i$. There exists a positive, monotonically decreasing (perhaps to 0 in the limit) function $\varepsilon_0^{P^d}(\cdot)$, such that for $J_1 \in \widetilde{\mathcal{J}}$, $J_2 \notin \widetilde{\mathcal{J}}$ with $|J_2| = |J_1| + 1$, we have

$$\mathcal{R}^*_{L,P^d,\mathcal{F}^{J_2}} \ge \mathcal{R}^*_{L,P^d,\mathcal{F}^{J_1}} + \varepsilon_0^{P^d}(d - |J_1|). \tag{9}$$

As mentioned, given the nature of the problem P^d, $\varepsilon_0^{P^d}(\cdot)$ can go to zero in the limit, when $d \to \infty$. d_0 can potentially grow with n as well, but we will restrict to the case when $d - d_0$ necessarily grows with n, for example, when d grows with n and $d_0 = O(d^\alpha)$ for $0 < \alpha < 1$.

REMARK 18. Recall $\delta_n^{P^d}(\cdot)$ from Theorem 2. Note that there are two different asymptotic conditions acting on $\delta_n^{P^d}(\cdot)$ with $\delta_n^{P^d}(\cdot) \to \varepsilon_0^{P^d}(\cdot)$ as $n \to \infty$ by Theorem 2. Now from above we also see that $\delta_n^{P^d}(d - d_0) \to 0$ as d and n go to infinity and as $\varepsilon_0^{P^d}(d - d_0)$ goes to zero in the limit.

In the next section, we will formally establish the main results which were discussed in Section 2.6.

4. Theoretical results.

In this section, we will prove Theorems 1A and 2A. To establish these theorems, we need a few additional results, most of which are provided in the Supplementary Material (Section S3). Here, we have summarized these results in two lemmas, one for each of the two settings, that we provide before going into the proofs.

LEMMA 4. *Assume the conditions of Theorem* 1A *in Section* 2.6. *Then*, *for a sequence* $\tau = o(n^{\frac{2\beta}{2\beta+1}})$ *with* $\tau \to \infty$, *the following statements hold*:

(i) *For* $J_1, J_2 \in \widetilde{\mathcal{J}}$ *such that* $J_1 \subseteq J_2 \subseteq J_*$, *there is a positive sequence* $\{\varepsilon_n\}$ *with* $\varepsilon_n \to 0$ *for which we have with* P^n *probability greater than* $1 - 2e^{-\tau}$,

$$\mathcal{R}^{\mathrm{reg},\lambda_n}_{L,D,H^{J_2}}(f_{D,\lambda_n,H^{J_2}}) \le \mathcal{R}^{\mathrm{reg},\lambda_n}_{L,D,H^{J_1}}(f_{D,\lambda_n,H^{J_1}}) + \varepsilon_n.$$

(ii) *For* $J_1 \in \widetilde{\mathcal{J}}$, $J_2 \notin \widetilde{\mathcal{J}}$ *with* $J_1 \subset J_2$, *and for the same sequence* $\{\varepsilon_n\}$ *defined above in* (i), *we have with* P^n *probability greater than* $1 - 2e^{-\tau}$,

$$\mathcal{R}^{\mathrm{reg},\lambda_n}_{L,D}(f_{D,\lambda_n,H^{J_2}}) \ge \mathcal{R}^{\mathrm{reg},\lambda_n}_{L,D}(f_{D,\lambda_n,H^{J_1}}) + \varepsilon_0 - \varepsilon_n.$$

(iii) ORACLE PROPERTY FOR RISK-RFE IN KM: *The infinite-sampled regularized risk for the empirical solution* f_{D,λ_n,H^J}, $\mathcal{R}^{\mathrm{reg},\lambda_n}_{L,P,H^J}(f_{D,\lambda_n,H^J})$ *converges in measure to* $\mathcal{R}^*_{L,P,H}$ [*and hence to* $\mathcal{R}^*_{L,P}$ *if the RKHS* H *is dense in* $\mathcal{L}_\infty(\mathcal{X})$] *iff* $J \in \widetilde{\mathcal{J}}$.

The proof of Lemma 4 is given in the Supplementary Material (Section S8.2). We are now ready to prove Theorem 1A.

4.1. *Proof of Theorem* 1A (*from Section* 3.3.3).

PROOF. (CS1) Let $\mathcal{X}^{J_*}$ be the correct input space and J_* be the correct set of dimensions to be removed with $|J_*| = d - d_0$. To prove the first part of Theorem 1A, 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 $\mathcal{J}^\circ$ be that correct path and $\mathcal{J}^\circ = \{J_0^\circ \equiv \{\cdot\}, J_1^\circ, \dots, J_{d-d_0}^\circ \equiv J_*\}$.

For notational ease, let us further define (we will use these later as well),

$$
\begin{aligned}
&\mathcal{RR}_{Q_2}^{Q_1}(J_1, J_2) := \mathcal{R}_{L,Q_2,H^{J_1}}^{\mathrm{reg},\lambda_n}(f_{Q_1,\lambda_n,H^{J_1}}) - \mathcal{R}_{L,Q_2,H^{J_2}}^{\mathrm{reg},\lambda_n}(f_{Q_1,\lambda_n,H^{J_2}}),\\
&\mathcal{RR}_{Q_2}^{Q_1}(J) := \mathcal{R}_{L,Q_2,H^{J}}^{\mathrm{reg},\lambda_n}(f_{Q_1,\lambda_n,H^{J}}) - \mathcal{R}^*_{L,P,H^J}.
\end{aligned}
\tag{10}
$$

From equation (3) in the proof of (i) in the Supplementary Material (Section S8.2), we have $\mathcal{RR}_D^D(J_{i+1}^\circ, J_i^\circ) \le \varepsilon_n$ with probability at least $1 - 2e^{-\tau}$, for $\varepsilon_n = (2c + 24\sqrt{2\tau} + 48K_2a^{2p})n^{-\frac{\beta}{2\beta+1}} + 40\tau n^{-\frac{4\beta+1}{2(2\beta+1)}}$. Now if $J_{i+1} \neq J_{i+1}^\circ$ be any other J such that $J_i^\circ \subset J_{i+1}$ with $|J_{i+1}| = |J_i^\circ| + 1$, then we have from equation (4) in the Supplementary Material (Section S8.2) that $\mathcal{RR}_D^D(J_{i+1}, J_i^\circ) > \varepsilon_0 - \varepsilon_n$ with probability at least $1 - 2e^{-\tau}$. Now as ε_0 is a fixed constant, $\exists$ a finite $N_{\varepsilon_0} > 0$ such that $\forall n \ge N_{\varepsilon_0}$, $2\varepsilon_n \le \varepsilon_0$. Without loss of generality, assuming $n \ge N_{\varepsilon_0}$, note that we also have $\mathcal{RR}_D^D(J_{i+1}^\circ, J_i^\circ) \le \varepsilon_n \le \varepsilon_0 - \varepsilon_n$ with probability at least $1 - 2e^{-\tau}$. Now if we choose $\tau = o(n^{\frac{2\beta}{2\beta+1}})$ with $\tau \to \infty$, then ε_n is such that $\varepsilon_n n^{\frac{\beta}{2\beta+1}} \to C$ for some $C \in (0, \infty)$. Hence for $\delta_n := \varepsilon_0 - \varepsilon_n$ with ε_n as defined above, we have $\mathcal{RR}_D^D(J_{i+1}, J_i^\circ) > \delta_n$, and $\mathcal{RR}_D^D(J_{i+1}^\circ, J_i^\circ) \le \delta_n$ with probability at least $1 - 2e^{-\tau}$. Then

$$
\begin{aligned}
&P(\textit{“risk-RFE finds the correct space”})\\
&\quad \ge P(\textit{“risk-RFE follows the path } \mathcal{J}^\circ \textit{ to the correct dimension space”})\\
&\quad = P(J_0 := J_0^\circ, J_1 := J_1^\circ, \dots, J_{d-d_0} := J_{d-d_0}^\circ, J_{d-d_0+1} := \varnothing)\\
&\quad = P(J_0 := J_0^\circ) P(J_1 := J_1^\circ | J_0^\circ) \cdots P(J_{d-d_0} := J_{d-d_0}^\circ | J_0^\circ, \dots, J_{d-d_0-1}^\circ)\\
&\qquad \times P(J_{d-d_0+1} := \varnothing | J_0^\circ, \dots, J_{d-d_0}^\circ),
\end{aligned}
$$

where “$J_{d-d_0+1} := \varnothing$” means the algorithm stops at that step. Note that $P(J_0 := J_0^\circ) = 1$ and then observe

$$
\begin{aligned}
&P(J_{i+1} := J_{i+1}^\circ | J_0^\circ, \dots, J_i^\circ)\\
&\quad = P(J_{i+1} := J_{i+1}^\circ | J_i^\circ) \qquad (\because \{J_0^\circ, \dots, J_{i-1}^\circ\} \textit{ have already been removed})\\
&\quad = P(\mathcal{RR}_D^D(J_{i+1}^\circ, J_i^\circ) \le \delta_n,\\
&\qquad \mathcal{RR}_D^D(J_{i+1}^\circ, J_i^\circ) < \mathcal{RR}_D^D(J_{i+1}^\bullet, J_i^\circ)\ \forall J_{i+1}^\bullet \neq J_{i+1}^\circ)\\
&\quad \ge P(\mathcal{RR}_D^D(J_{i+1}^\circ, J_i^\circ) \le \delta_n, \delta_n < \mathcal{RR}_D^D(J_{i+1}^\bullet, J_i^\circ)\ \forall J_{i+1}^\bullet \neq J_{i+1}^\circ)\\
&\quad \ge 1 - P(\mathcal{RR}_D^D(J_{i+1}^\circ, J_i^\circ) > \delta_n) - \sum_{J_{i+1}^\bullet \neq J_{i+1}^\circ} P(\mathcal{RR}_D^D(J_{i+1}^\bullet, J_i^\circ) \le \delta_n)\\
&\quad \ge 1 - 2e^{-\tau} - 2(d - i - 1)e^{-\tau}\\
&\quad = 1 - 2(d - i)e^{-\tau}.
\end{aligned}
$$

Also note that

$$\begin{aligned}&P\bigl(J_{d-d_0+1} := \varnothing | J_0^\circ, \dots, J_{d-d_0}^\circ\bigr)\\ &\quad = P\bigl(\mathcal{RR}_D^D\bigl(J_{d-d_0+1}, J_{d-d_0}^\circ\bigr) > \delta_n \ \forall J_{d-d_0+1} \supseteq J_{d-d_0}^\circ\bigr) \ge 1 - 2d_0 e^{-\tau}.\end{aligned}$$

Hence $P(\textit{“risk-RFE finds the correct space”}) \ge \prod_{i=0}^{d-d_0}(1-2(d-i)e^{-\tau})$. Now for $\tau = o(n^{\frac{2\beta}{2\beta+1}})$ with $\tau \to \infty$, $P(\textit{“risk-RFE finds the correct space”}) \to 1$ 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 $\mathcal{J}_1, \dots, \mathcal{J}_N$ be an enumeration of all possible such paths. Define “C-set” for a given set J_i (where index i denotes the ith iteration of risk-RFE) as $\mathcal{C}(J_i) := \{J_{i+1} : J_i, J_{i+1} \in \mathcal{J}_k \text{ for some } k\}$. Now,

$$\begin{aligned}&P\bigl(\textit{“risk-RFE finds the correct space”}\bigr)\\ &\quad \ge P\bigl(J_0 := J_0^\circ, J_1 := J_1^\circ \in \mathcal{C}\bigl(J_0^\circ\bigr), \dots, J_{d-d_0+1} := \varnothing\bigr)\\ &\quad = P\bigl(J_0 := J_0^\circ\bigr) P\bigl(J_1 := J_1^\circ \in \mathcal{C}\bigl(J_0^\circ\bigr) | J_0^\circ\bigr) \cdots P\bigl(J_{d-d_0+1} := \varnothing | J_{d-d_0}^\circ\bigr).\end{aligned}$$

Again as before $P(J_0 := J_0^\circ) = 1$ and $P(J_{d-d_0+1} := \varnothing | J_{d-d_0}^\circ) \ge 1 - 2d_0 e^{-\tau}$. Now note

$$\begin{aligned}&P\bigl(J_{i+1} := J_{i+1}^\circ \in \mathcal{C}\bigl(J_i^\circ\bigr) | J_i^\circ\bigr)\\ &\quad \ge P\bigl(\mathcal{RR}_D^D\bigl(J_{i+1}^\circ, J_i^\circ\bigr) \le \delta_n \ \forall J_{i+1}^\circ \in \mathcal{C}\bigl(J_i^\circ\bigr),\\ &\qquad \delta_n < \mathcal{RR}_D^D\bigl(J_{i+1}^\bullet, J_i^\circ\bigr) \ \forall J_{i+1}^\bullet \notin \mathcal{C}\bigl(J_i^\circ\bigr)\bigr)\\ &\quad \ge 1 - \sum_{J_{i+1}^\circ \in \mathcal{C}(J_i^\circ)} P\bigl(\mathcal{RR}_D^D\bigl(J_{i+1}^\circ, J_i^\circ\bigr) > \delta_n\bigr)\\ &\qquad - \sum_{J_{i+1}^\bullet \notin \mathcal{C}(J_i^\circ)} P\bigl(\mathcal{RR}_D^D\bigl(J_{i+1}^\bullet, J_i^\circ\bigr) \le \delta_n\bigr)\\ &\quad \ge 1 - 2\bigl|\mathcal{C}\bigl(J_i^\circ\bigr)\bigr| e^{-\tau} - 2\bigl|\mathcal{C}\bigl(J_i^\circ\bigr)^c\bigr| e^{-\tau} = 1 - 2(d-i)e^{-\tau},\end{aligned}$$

since $|\mathcal{C}(J_i^\circ)| + |\mathcal{C}(J_i^\circ)^c| = d - i$. Hence again we have that

$$P\bigl(\textit{“risk-RFE finds the correct space”}\bigr) \ge \prod_{i=0}^{d-d_0}\bigl(1 - 2(d-i)e^{-\tau}\bigr).$$

Now for $\tau = o(n^{\frac{2\beta}{2\beta+1}})$ with $\tau \to \infty$, $P(\textit{“risk-RFE finds the correct space”}) \to 1$ as $n \to \infty$.

(CS2) To prove the second part of Theorem 1A, suppose that J_{end} is the last iteration of the algorithm in risk-RFE. Then using (5) in the Supplementary Material

(Section S8.2), and observing that $\eta_n < \varepsilon_n < \delta_n$ $\forall n \geq N_{\varepsilon_0}$, repeating arguments given at the beginning of the first part of the proof we have that

$$\begin{aligned}
&P\big(\big|\mathcal{R}^{\mathrm{reg},\lambda_n}_{L,P,H^{J_{\mathrm{end}}}}(f_{D,\lambda_n,H^{J_{\mathrm{end}}}}) - \mathcal{R}^*_{L,P,H}\big| \leq \delta_n\big)\\
&\quad = P\big(\big|\mathcal{R}^{\mathrm{reg},\lambda_n}_{L,P,H^{J_*}}(f_{D,\lambda_n,H^{J_*}}) - \mathcal{R}^*_{L,P,H}\big| \leq \delta_n\big)P(J_{\mathrm{end}} = J_*)\\
&\qquad + P\big(\big|\mathcal{R}^{\mathrm{reg},\lambda_n}_{L,P,H^{J_*}}(f_{D,\lambda_n,H^{J_*}}) - \mathcal{R}^*_{L,P,H}\big| \leq \delta_n | J_{\mathrm{end}} \neq J_*\big)P(J_{\mathrm{end}} \neq J_*)\\
&\quad \geq P\big(\big|\mathcal{R}^{\mathrm{reg},\lambda_n}_{L,P,H^{J_*}}(f_{D,\lambda_n,H^{J_*}}) - \mathcal{R}^*_{L,P,H}\big| \leq \delta_n\big)P(J_{\mathrm{end}} = J_*)\\
&\quad \geq \big(1 - e^{-\tau}\big)\prod_{i=0}^{d_0}\big(1 - 2(d-i)e^{-\tau}\big).
\end{aligned}$$

So for $\tau = o(n^{\frac{2\beta}{2\beta+1}})$ with $\tau \to \infty$,

$$P\big(\big|\mathcal{R}^{\mathrm{reg},\lambda_n}_{L,P,H^{J_{\mathrm{end}}}}(f_{D,\lambda_n,H^{J_{\mathrm{end}}}}) - \mathcal{R}^*_{L,P,H}\big| \leq \delta_n\big) \to 1$$

with $n \to \infty$. □

4.2. *Proof of Theorem* 2A (*from Section* 3.3.3). We only note the modifications that are required in the above proof to establish Theorem 2A. But first we formally define the regularity conditions (RC2) below:

(RC2) There exist constants $\tilde{a} \geq 1$ and some $p \in (0, 1)$ such that for $i \geq 1$, $\mathbb{E}_{D_{\mathcal{X}} \sim P^{d,n}_{\mathcal{X}}} e_i(id : H \mapsto L_\infty(D_{\mathcal{X}})) \leq f(d)\tilde{a}i^{-\frac{1}{2p}}$, and there exists a $\tilde{c} > 0$ and $\beta \in (0, 1]$ such that $A^{J_*}_2(\lambda) \leq g(d_0)\tilde{c}\lambda^\beta$ (for J_* and d_0 defined before), for all $\lambda \geq 0$, and for functions $f(\cdot)$, $g(\cdot)$ on $\mathbb{N} \mapsto \mathbb{R}$. We also assume that there exist $\gamma_1, \gamma_2 > 0$ with $\max(2\gamma_1, \gamma_2) \leq \frac{\beta}{2\beta+1}$, such that (i) $f(d) = O(n^{\frac{\beta}{2(2\beta+1)} - \gamma_1})$, (ii) $g(d_0) = O(n^{\frac{\beta}{2\beta+1} - \gamma_2})$, and (iii) $d = o(e^{0.5n^{\frac{2\beta}{2\beta+1}}})$.

Under condition (RC2), it can be seen that the modifications required for the bounds given in Lemma S4–Corollary S7 from the Supplementary Material (Section S3) can be achieved by replacing a by $f(d)\tilde{a}$ and c by $g(d_0)\tilde{c}$. Lemma 4 can now be restated as Lemma 4* below, by replacing ε_n by $\varepsilon_{n,d} = (2\tilde{c}g(d_0) + 24\sqrt{2\tau} + 48K_2\tilde{a}^{2p}f(d)^{2p})n^{-\frac{\beta}{2\beta+1}} + 40\tau n^{-\frac{4\beta+1}{2(2\beta+1)}}$.

LEMMA 4*. *Assume the conditions of Theorem* 2A *in Section* 2.6. *Then for a sequence* $\tau = o(n^{\frac{2\beta}{2\beta+1}})$, *the following statements continue to hold*:

(i) *For* $J_1, J_2 \in \tilde{\mathcal{J}}$ *such that* $J_1 \subseteq J_2 \subseteq J_*$ *and a positive sequence* $\{\varepsilon_{n,d}\}$ *with* $\varepsilon_{n,d} \to 0$, $\mathcal{R}^{\mathrm{reg},\lambda_n}_{L,D,H^{J_2}}(f_{D,\lambda_n,H^{J_2}}) \leq \mathcal{R}^{\mathrm{reg},\lambda_n}_{L,D,H^{J_1}}(f_{D,\lambda_n,H^{J_1}}) + \varepsilon_{n,d}$ *occurs with* $P^{d,n}$ *greater than* $1 - 2e^{-\tau}$.

(ii) *For* $J_1 \in \widetilde{\mathcal{J}}$, $J_2 \notin \widetilde{\mathcal{J}}$ *with* $J_1 \subset J_2$, *and for the same sequence* $\{\varepsilon_{n,d}\}$ *defined above in* (i), $\mathcal{R}^{\mathrm{reg},\lambda_n}_{L,D,H^{J_2}}(f_{D,\lambda_n,H^{J_2}}) \geq \mathcal{R}^{\mathrm{reg},\lambda_n}_{L,D,H^{J_1}}(f_{D,\lambda_n,H^{J_1}}) + \varepsilon_0^{P^d}(d - |J_1|) - \varepsilon_{n,d}$ *occurs with* $P^{d,n}$ *probability greater than* $1 - 2e^{-\tau}$.

(iii) ORACLE PROPERTY FOR RISK-RFE: *Continues to hold as before.*

Under the modified statements, consistency can be established. It can be easily observed that the initial steps in the proof of Theorem 1A in Section 4.1 continue to hold by taking $\delta_n^{P^d}(d - |J|) = \varepsilon_0^{P^d}(d - |J|) - \varepsilon_{n,d}$ for design $\mathcal{X}^J$, and now we can further assume that $\sup_{d \in \mathbb{N}} \liminf_{n \to \infty} \frac{\varepsilon_0^{P^d}(d-d_0)}{\varepsilon_{n,d}} > 2$. This allows us to define a sequence $\{N_1, \ldots, N_d, \ldots\}$, such that $2\varepsilon_{n,d} \leq \varepsilon_0^{P^d}(d - d_0)$, whenever $n > N_d$. Since $\varepsilon_0^{P^d}(\cdot)$ is a decreasing function, the subsequent steps follow and we arrive at

$$
\begin{aligned}
P(\text{“\textit{RFE finds the correct space}”}) &\geq \prod_{i=0}^{d-d_0} (1 - 2(d-i)e^{-\tau}) \\
&\gtrsim (1 - 2de^{-\tau})^d, \qquad (11)
\end{aligned}
$$

where the last approximate inequality follows if we can ensure that $2de^{-\tau} < 1$ for sufficiently large n, d. Also see that the above implies

$$
(1 - 2de^{-\tau})^d = \left(\left(1 - \frac{2d}{e^{\tau}}\right)^{-\frac{e^{\tau}}{2d}}\right)^{-\frac{2d^2}{e^{\tau}}}. \qquad (12)
$$

Thus if we require $d^2 e^{-\tau} \to 0$ when $n, d \to \infty$, then (11) is satisfied, and (12) converges to 1. Consequently, for consistency results to hold, d needs to grow slower than a certain rate in terms of the sample size n. Since τ can be chosen to be $o(n^{\frac{2\beta}{2\beta+1}})$ with $\tau \to \infty$, it implies that $de^{-\tau/2} \approx de^{-0.5n^{\frac{2\beta}{2\beta+1}}}$, and hence $d = o(e^{0.5n^{\frac{2\beta}{2\beta+1}}})$ suffices. We now need to ensure that asymptotically $\varepsilon_{n,d}$ goes to 0. Since we can let $\tau = o(n^{\frac{2\beta}{2\beta+1}})$, and p is a constant in $(0, 1)$ [from Proposition S5 in the Supplementary Material (Section S3)], this forces $\varepsilon_{n,d}$ to satisfy $\varepsilon_{n,d} \leq c_1 g(d_0) n^{-\frac{\beta}{2\beta+1}} + c_2 f(d)^2 n^{-\frac{\beta}{2\beta+1}} + o(1)$. Now using conditions on $f(d)$ and $g(d_0)$ from (RC2), it can be seen that $\varepsilon_{n,d} < \tilde{c_1} n^{-\gamma_2} + \tilde{c_2} n^{-2\gamma_1}$. Hence for $\gamma = \min(2\gamma_1, \gamma_2)$, $\varepsilon_{n,d}$ satisfies the condition that $\varepsilon_{n,d} n^{\gamma} \to K$ as $n, d \to \infty$, for some $K \in [0, \infty)$.

REMARK 19. The functional bounds $f(\cdot)$ and $g(\cdot)$ are characteristics of the data generating mechanism of the input $\mathcal{X}$ and the output Y through $P^{d,n}$, the RKHS H used for optimization, and the loss function L. Thus for a given problem, we have a specific representation of the functions $f(\cdot)$ and $g(\cdot)$, and restrictions on the dimensionality growth for risk-RFE is obtained by making sure conditions on f, g and d given in (RC2) hold. Let us now look at the allowed dimensionality growth under some special forms of $f(\cdot)$ and $g(\cdot)$. Note that (RC2C) summarizes the allowed dimensionality growth in a typical kernel machines classification framework with the Gaussian RBF kernel.

(RC2A) $f(d) = c_1$ and $g(d_0) = c_2$. Under this setting, we can allow rates as high as $d = o(e^{0.5n^{\frac{2\beta}{2\beta+1}}})$ with $d_0 = O(d^\alpha) < d$ where $0 < \alpha < 1$, and the algorithm will continue to be consistent for $\gamma = \frac{\beta}{2\beta+1}$.

(RC2B) $f(d) = e^d$ and $g(d_0) = e^{d_0}$. Under this setting, it can be seen that we need $d = O(\log n)$. We can still allow $d_0 = O(d^\alpha) < d$ with $0 < \alpha < 1$, and the algorithm is consistent for $0 < \gamma < \frac{\beta}{2\beta+1}$.

(RC2C) Dimensionality growth in the Gaussian RBF kernel: $f(d) = e^d$ and $g(d) = d_0^{cd_0}$. Under this, we can continue to have $d = O(\log n)$ but now we need $d_0 \log d_0 = O(\log n)$ with $d_0 < d$, and the algorithm is consistent for $0 < \gamma < \frac{\beta}{2\beta+1}$.

5. Case studies.

In this section, we show the validity of our results in some known settings of learning through risk minimization. We discuss two pertinent examples here, and two more case studies (simple linear regression and protein classification) are discussed in the Supplementary Material (Section S4).

5.1. *Case study* 1: *Kernel machines with a Gaussian RBF kernel.*

Here, we study the application of risk-RFE in the classic KM 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 decision rule is the Bayes' function $f^*_{L,P}$ for the classification loss $L_{\text{BC}}(x, y, f(x)) = 1\{y \cdot \text{sign}(f(x)) \neq 1\}$. In practice, since L_{BC} is nonconvex, it is usually replaced by the hinge loss function $L_{\text{HL}}(x, y, f(x)) = \max\{0, 1 - yf(x)\}$. For KMs 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 $H_\gamma(\mathcal{X})$ with kernel $k_\gamma(x, y) = e^{-\gamma^2\|x-y\|_2^2}$, where $\sigma := 1/\gamma$ is the width of the kernel k_γ.

Lemma 5. *For classification using kernel machines with a Gaussian RBF kernel, the risk-RFE defined for $\delta = \varepsilon_0 - \mathcal{R}(n)$ where $\mathcal{R}(n)^{-1}$ is $O(n^{\frac{\beta}{2\beta+1}})$, with $\beta = \frac{\alpha}{\alpha+1}$ and $0 < \alpha < \infty$ being the geometric noise exponent of P on $\mathcal{X} \times \{-1, 1\}$.*[4]

[4] For a discussion on the geometric noise exponent, we refer our readers to Steinwart and Scovel (2007).

Lemma 5 gives us a precise characterization of δ_n in this setting, in terms of the geometric noise exponent of P on $\mathcal{X} \times \{1, -1\}$. In the proof of Lemma 5, which is given in the Supplementary Material (Section S8.3), we show that the assumptions and conditions required for consistency of the risk-RFE algorithm in classification with the Gaussian RBF kernel are properly satisfied.

5.2. *Case study* 2: *Image classification with* χ^2 *kernel.* Using color histograms as an image representation technique is a useful tool in indexing or retrieving images because of the reasonable performance that can be obtained in spite of their extreme simplicity [see Swain and Ballard (1992)]. Image classification using histogram representation has become a popular option in such settings, and the kernel machines approach is considered a good classification technique here [see Chapelle, Haffner and Vapnik (1999)].

Selecting the kernel is important in these problems, and generalized RBF kernels of the form $K_\rho^{d\text{-RBF}}(x, y) = e^{-\rho d(x,y)}$ have proven useful for classification here. When the inputs are images, the histograms produced generate discrete densities and suitable comparison functions such as the χ^2 function are preferred over the L_2 norm that generates the usual Gaussian RBF kernel and have been used extensively for histogram comparisons [Schiele and Crowley (1996)]. The χ^2 distance is given as $d_{\chi^2}(x, y) = \sum_i \frac{(x_i - y_i)^2}{x_i + y_i}$, and hence the χ^2 kernel has the form

$$K_\rho^{\chi^2\text{-RBF}}(x, y) = e^{-\rho \sum_i \frac{(x_i - y_i)^2}{x_i + y_i}}.$$

To establish consistency for risk-RFE here, we need to verify regularity conditions (RC0) and (RC1) given in Section 2.6. Note that we have already established the conditions for hinge loss L_{HL} in the case study 5.1. The $K_\rho^{\chi^2\text{-RBF}}$ kernel is continuous, and the input space is separable; hence separability of $H_\rho^{\chi^2\text{-RBF}}$ follows from Lemma 4.33 of SC08. It also follows that $\|K_\rho^{\chi^2\text{-RBF}}\|_\infty \leq 1$. Since the kernel $K_\rho^{\chi^2\text{-RBF}}$ is infinitely many times differentiable (such as the standard RBF kernel), Theorem 6.26 along with arguments in Theorem 7.34 with Corollary 7.31 of SC08, can give us an explicit formulation for the average entropy of this RKHS class, which is very similar to the ones that we saw in the Supplementary Material (Section S8.3), with $a := c_{\varepsilon,p}\rho^{\frac{(1-p)(1+\varepsilon)d}{4p}}$ for all $\rho \geq 1$, for all $\varepsilon > 0$, $d/(d+\tau) < p < 1$, and constant $c_{\varepsilon,p}$ which depends only on p and a given ε, and where $\tau \in (0, \infty]$ is the tail exponent of the distribution $P_{\mathcal{X}}$ (see Chapter 7 of SC08, for definition of the tail exponent of a distribution). Hence consistency follows.

6. Simulation studies. Now we illustrate the usefulness of risk-RFE for feature elimination in KMs through a simulation study. The first key step is to formulate the stopping rule for risk-RFE in a practical setting, thus we start off with a discussion of how to intelligently select the subset of features using the risk-RFE algorithm for a given problem below. We also evaluate the performance of risk-RFE in an extensive simulation exercise incorporating various linear and nonlinear settings, with focus mainly on the nonlinear setup. For brevity, however, the details of these simulations and a comprehensive discussion of our findings are provided in the Supplementary Material (Section S5). Nonetheless, we will briefly discuss our findings from our simulations incorporating the nonlinear settings here.

6.1. *Selection of features.* One crucial question we face in feature elimination is when to stop. The original RFE algorithm can only output the ranked features, so one crucial aspect of risk-RFE is that it can be used not only to rank features, but also for automatic selection of the optimal subset. This is seen by noting that in our theory, we proposed the existence of a gap ε_0, and showed that asymptotically the empirical regularized risk of a model with at least one important feature missing exceeds that of a correct model by at least this amount. Practically, it is very difficult to characterize this gap for a given setting from the theory directly, but its existence can be observed from plotting the objective function values at each iteration for the entire set of features. Here, we use a "Scree graph" of the objective function (see Figure 2) to build an auto-selection rule [see Chapter 6 of Jolliffe (2002) for scree graphs].

Looking at Figure 2, which plots the objective function values obtained at successive runs of the algorithm for a given setting, it seems plausible that a change-point model can be used for curve-fitting here, as we expect a change in the slope of the objective function as soon as we start eliminating signals from the model (because of the aforementioned gap). Thus, one can fit a change-point regression model on the empirical values of the objective function and infer the estimated change point as the ad hoc stopping rule. Different trends (linear, quadratic, etc.)

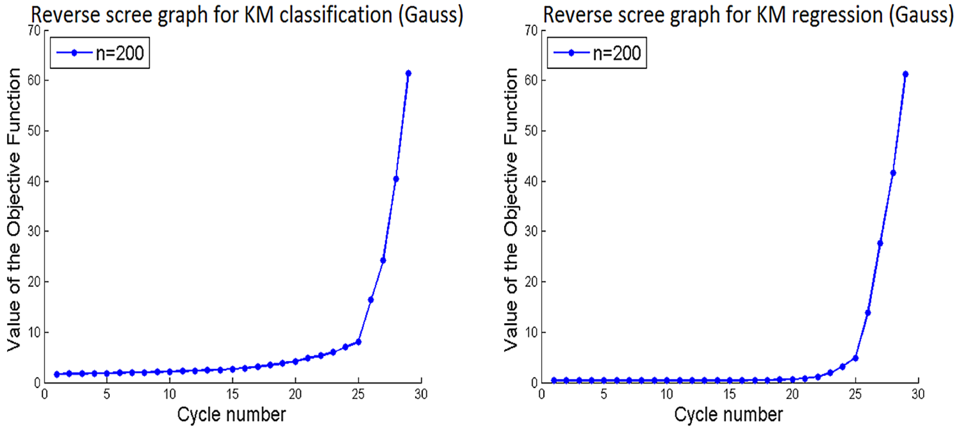

FIG. 2. *Reverse scree graph of the objective function values for one simulation run in* $d = 30$, $d_0 = 5$, *with* (a) *KM classification with Gaussian kernel and* (b) *KM regression with Gaussian kernel.*

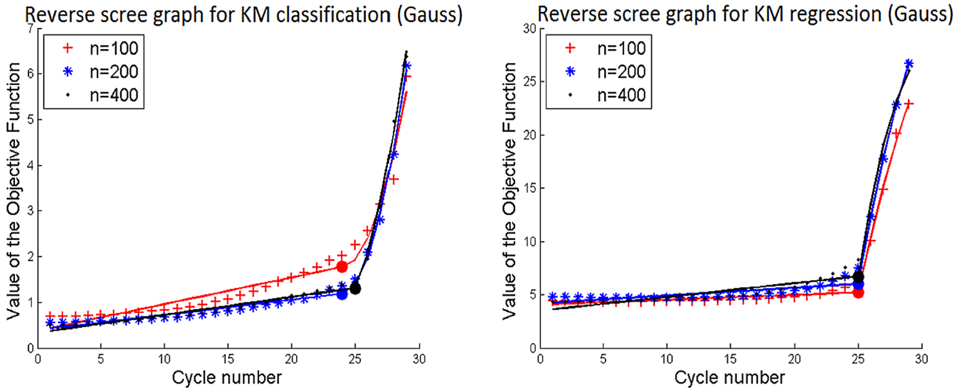

FIG. 3. *Linear-quadratic mixture change-point analysis when $d = 30$, $d_0 = 5$, for* (a) *KM classification with Gaussian kernel for comparable cross validation values of λ and kernel width γ and* (b) *KM regression with Gaussian kernel for comparable cross validation values of λ for varying sample sizes. The bold dots represent the estimated change points.*

can be fit on each side of the change point, depending on how the values for the objective function are scattered on the graph. For our simulation examples, a mixture of linear trend on the left and quadratic trend on the right seemed to work best. Figure 3 shows our analysis using the mixture of linear-quadratic fits.

6.2. *Performance of risk-RFE in the nonlinear space* (*results*). In the Supplementary Material (Section S5), we look at some nonlinear settings in both classification and regression to ascertain the performance of risk-RFE when the underlying functional form of the decision function is nonlinear. As a first confirmatory step, we compare the performance of risk-RFE (RRFE) with the original RFE (GRFE) proposed by Guyon et al. (2002). We also compare it with linear selection methods like SCAD-SVM (linear KM with SCAD) and logistic regression with LASSO (Log Reg Lasso) in classification, and with linear regression with LASSO (Lin Reg Lasso) in regression, to show the necessity of considering nonlinear feature selection methods when the underlying relationship in nonlinear. For risk-RFE, we consider feature selection through the proposed change-point model (CP), and also through a naive ranks approach (NR), that only considers the first d_0 highest ranked signals, using the oracle knowledge of the true number of signals d_0 in each setting. This is done to make it comparable to GRFE, which does not have an inherent subset selection rule. Table S5.1 compiles results from the different nonlinear simulation settings in classification, while Table S5.2 compiles those from regression, both under the presence and absence of colinearity. Here, we summarize our findings from that analysis:

(1) The risk-RFE procedures (both RRFE-NR with naive ranks, and RRFE-CP with the change-point model) dominate performance in choosing the correct

features consistently, supporting our asymptotic claims for consistency of the algorithm.

(2) GRFE-NR does relatively well in classification, although never better than RRFE-NR, and this dominance increases with dimension size. In regression, however, its performance is quite poor, and in higher dimensions, it struggles to find any signals. The test set performance is consistently better for RRFE-NR than for GRFE-NR both in classification and regression, unless they perform equally well.

(3) In classification, the RRFE-CP procedure yields higher misclassification error rates than both the NR procedures, owing to the fact that the NR methods depend on knowing the oracle number of signals beforehand.

(4) In regression, RRFE-NR dominates RRFE-CP in test error rates, but the RRFE-CP procedure dominates GGRE-NR in that metric more often than not, especially when the latter struggles to find features.

(5) As expected, the linear selection methods like SVM-SCAD and logistic regression with LASSO in classification and linear regression with LASSO in regression perform very poorly in most situations, both in terms of feature selection, and in terms of misclassification or average test error rates.

(6) The test set performances of the standard KM procedure without selection is consistently worse than those obtained after using the RFE procedures. Using a linear feature selection method often yields test set performance which is even worse than the standard nonlinear KM procedure without selection, showing the usefulness of nonlinear feature selection methods like risk-RFE.

(7) In classification risk-RFE performs equally well when we have colinearity versus when we do not. In regression, it tends to do better when there is no colinearity, and this effect is most pronounced when dimension (and signal) sizes increase.

Overall we can conclude that in moderately high-dimensional classification and regression, when there is enough suspicion that the underlying structure is not linear, using risk-RFE is a very safe option.

7. Example applications.

We apply the risk-RFE algorithm to three feature selection applications: feature selection in vowel recognition data, feature selection in predicting total UPDRS (unified Parkinson's disease rating scale) scores in people with early-stage Parkinson's disease and feature selection in predicting "Per Capita Violent Crimes" in the Crimes dataset. We briefly discuss our results here, while detailed discussions on each dataset and results for each analysis, are given in the Supplementary Material (Section S7).

- In the classification analysis with *vowel recognition data*, risk-RFE only chooses 4 out of 10 features, and is able to achieve a more than 40% drop in misclassification error than its nearest competitor (see Table 1).
- In the regression analysis with *Parkinson's data*, risk-RFE algorithm only chooses 3 out of 20 features, and achieves a 13% drop in test error rates from when we use all of the 20 features (see Table 2).

TABLE 1
Results from the data analyses (classification)

Vowel data		
Method	**Mean test error**	**Ave. no of features**
KM wRisk-RFE	0.08	4
KM woSEL	0.18	10
SCAD SVM	0.14	8
Log Reg wLASSO	0.19	9

- In the regression analysis with *Crimes dataset*, the risk-RFE algorithm does not show any improvement in test error rates over the competing methods, but it yields a final model which is much more parsimonious than any of the competing methods, choosing only 11 features out of 101 (see Table 2).

8. Discussion. We proposed an algorithm for feature elimination in kernel machines for moderately high dimensions. We studied its theoretical properties, and showed that it is consistent in finding the correct feature space, even when the size of the design matrix grows with the sample size. We discussed the natural assumptions required to enable this consistency, and showed that these are satisfied in many practical settings through the study of a few case studies where our method becomes readily applicable. We also provided a short simulation study to illustrate the method and discussed a practical way for choosing the correct subset of features. We established the existence of a gap in the rate of change of the objective function at the point where our feature elimination method starts removing the essential features of the learning problem. This motivated us to use a scree plot of the objective function values at each iteration, and indeed our simulation results support our approach by visually exhibiting this gap in the plots. Moreover, the graphical interpretation of the scree plot motivated the use of change-point regression to select the correct feature space. It would thus be interesting to conduct a

TABLE 2
Results from the data analyses (regression)

	Parkinsons data		**Crimes data**	
Method	**Mean test error**	**Ave. no of features**	**Mean test error**	**Ave. no of features**
KM wRisk-RFE	10.08	3	0.02	10.8
KM woSEL	11.42	20	0.02	101
Lin Reg wLASSO	87.80	17.4	0.02	37.8

more detailed and formal analysis of this gap in real life settings to facilitate more efficient and automated practical solutions.

From our discussion in Section 4.2, we saw that when dimension d grows with n, risk-RFE is most effective when used under certain restrictions on the design size d relative to the sample size n. The theory for the consistency of the algorithm also accounted for a gradual growth in the signal size d_0 relative to the growth on n. However, as one of our reviewers pointed out, in real-life problems with truly high-dimensional covariate spaces, and/or problems with a relatively large number of signals, the risk-RFE algorithm may not be as scalable as some of the current state of the art ultrahigh-dimensional feature selection methods like the sure independence screening (SIS) method of Fan and Lv (2008) (or other SIS based methods such as DC-SIS [Li, Zhong and Zhu (2012)]), and thus risk-RFE needs to be used with some caution in problems which either have a very large covariate size or have the potential to contain a large number of signals. Typically, in linear models, methods like SIS are used to effectively screen models from ultra high dimensions to a lower dimensional setting, wherein more meaningful lower dimensional methods such as SCAD and LASSO become applicable. That is what we propose here as well, that is, when dealing with ultrahigh dimensions in KM, risk-RFE should be used in conjunction with one of these methods. As we saw, if the underlying model is nonlinear, using risk-RFE after initial feature screening would enhance the performance of the KM function compared to other available techniques.

To our knowledge, only very limited analysis has been done on the properties of variable selection algorithms under such general transformations of the input space as occur in kernel machines. Hence, the results generated in this paper are a good starting point for similar analyses in other settings. It would also be interesting to analyze risk-RFE in censored support vector regression [see Goldberg and Kosorok (2017)], other machine learning problems (including reinforcement learning), or other penalized risk minimization problems.

GLOSSARY

k: $k(x, y) = \langle \phi(x), \phi(x) \rangle_H$, the kernel function associated with a given reproducing kernel Hilbert space (RKHS).

$\mathcal{R}_{L,P}$: $\mathcal{R}_{L,P}(f) = E_P[L(X, Y, f(X))]$.

$\mathcal{L}_0$: The set of all measurable functions on a given space, that is, $\mathcal{L}_0(\mathcal{X}) = \{f : \mathcal{X} \mapsto \mathbb{R}, f \text{ is measurable}\}$.

$f_{P,\mathcal{F}}$: $f_{P,\mathcal{F}} = \arg\min_{f \in \mathcal{F}} \mathcal{R}_{L,P}(f)$ is the minimizer of infinite-sample risk within $\mathcal{F}$.

$\mathcal{R}^*$: The minimal risk attained within a space, for example, $\mathcal{R}^*_{L,P,\mathcal{F}} = \mathcal{R}_{L,P}(f_{P,\mathcal{F}})$.

L: $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \mapsto [0, \infty]$ is a convex, locally Lipschitz continuous and measurable loss function.

H: A separable RKHS of the measurable kernel k on $\mathcal{X}$.

$f_{P,\lambda,H}$: The infinite-sampled version of the Kernel Machines solution. For mathematical definition see (1).

$f_{D,\lambda,H}$: The empirical Kernel Machines solution. For mathematical definition see (2).

$\mathcal{R}_{L,D}$: $\mathcal{R}_{L,D}(f) = \frac{1}{n}\sum_{i=1}^{n} L(X_i, Y_i, f(X_i))$ for a given data D of size n.

reg: Implies regularized risk, for example $\mathcal{R}^{\text{reg},\lambda}_{L,Q,\mathcal{F}}(f) = \lambda\|f\|^2_{\mathcal{F}} + \mathcal{R}_{L,Q}(f)$.

π: The projection map. See Definition 1.

e_i: The ith entropy number for a given metric space, for mathematical definition see (5).

A_2: The approximation error for a given optimization space and a λ, for example, $A_2^J(\lambda) = \mathcal{R}^{\text{reg},\lambda}_{L,P,H^J}(f_{P,\lambda,H^J}) - \mathcal{R}^*_{L,P,H^J}$.

$\mathbb{E}_{D_{\mathcal{X}}\sim P^n_{\mathcal{X}}}$: The expectation w.r.t. $P^n_{\mathcal{X}}$ for data $D_{\mathcal{X}} \equiv \{\mathcal{X}_1, \dots, \mathcal{X}_n\}$ being i.i.d. copies of $\mathcal{X} \sim P_{\mathcal{X}}$. Similarly $\mathbb{E}_{D\sim P^n}$ for the joint measure P and full data D.

$L_\infty(D_{\mathcal{X}})$: The space of equivalence classes of all bounded measurable functions w.r.t. the measure $D_{\mathcal{X}}$.

$\mathcal{L}_\infty$: The set of all bounded measurable functions on a given space, that is, $\mathcal{L}_\infty(\mathcal{X}) = \{f : \mathcal{X} \mapsto \mathbb{R}, f \text{ measurable and } \|f\|_\infty < \infty\}$.

Acknowledgments. The authors are grateful to the anonymous reviewers, the Associate Editor and the Editor for their helpful suggestions and comments. 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.

SUPPLEMENTARY MATERIAL

Additional materials and Matlab codes (DOI: 10.1214/18-AOS1696SUPP; .pdf). Due to space constraint, all remaining proofs and additional details are provided in a separate file. Codes are available at http://www.bios.unc.edu/~kosorok/RFE.html.

REFERENCES

AKSU, Y. (2014). Fast SVM-based feature elimination utilizing data radius, hard-margin, soft-margin. Preprint. Available at arXiv:1210.4460v4.

AKSU, Y., MILLER, D. J., KESIDIS, G. and YANG, Q. X. (2010). Margin-maximizing feature elimination methods for linear and nonlinear kernel-based discriminant functions. *IEEE Trans. Neural Netw.* **21** 701–717.

ALLEN, G. I. (2013). Automatic feature selection via weighted kernels and regularization. *J. Comput. Graph. Statist.* **22** 284–299. MR3173715

BRADLEY, P. S. and MANGASARIAN, O. L. (1998). Feature selection via concave minimization and support vector machines. In *Machine Learning Proceedings of the Fifteenth International Conference* (*ICML* 1998) 82–90. Morgan Kaufmann, San Francisco, CA.

CHAPELLE, O., HAFFNER, P. and VAPNIK, V. N. (1999). Support vector machines for histogram-based image classification. *IEEE Trans. Neural Netw.* **10** 1055–1064.

DASGUPTA, S., GOLDBERG, Y. and KOSOROK, M. R (2019). Supplement to "Feature elimination in kernel machines in moderately high dimensions." DOI:10.1214/18-AOS1696SUPP.

FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 849–911. MR2530322

GOLDBERG, Y. and KOSOROK, M. R. (2017). Support vector regression for right censored data. *Electron. J. Stat.* **11** 532–569. MR3619316

GUYON, I., WESTON, J., BARNHILL, S. and VAPNIK, V. (2002). Gene selection for cancer classification using support vector machines. *Mach. Learn.* **46** 389–422.

HIDALGO-MUÑOZ, A. R., LÓPEZ, M. M., SANTOS, I. M., PEREIRA, A. T., VÁZQUEZ-MARRUFO, M., GALVAO-CARMONA, A. and TOMÉ, A. M. (2013). Application of SVM-RFE on EEG signals for detecting the most relevant scalp regions linked to affective valence processing. *Expert Syst. Appl.* **40** 2102–2108.

HU, X., SCHWARZ, J. K., LEWIS, J. S., HUETTNER, P. C., RADER, J. S., DEASY, J. O., GRIGSBY, P. W. and WANG, X. (2010). A microRNA expression signature for cervical cancer prognosis. *Cancer Res.* **70** 1441–1448.

JOLLIFFE, I. T. (2002). *Principal Component Analysis*, 2nd ed. Springer, New York. MR2036084

LESLIE, C. S., ESKIN, E., COHEN, A., WESTON, J. and NOBLE, W. S. (2004). Mismatch string kernels for discriminative protein classification. *Bioinformatics* **20** 467–476.

LI, R., ZHONG, W. and ZHU, L. (2012). Feature screening via distance correlation learning. *J. Amer. Statist. Assoc.* **107** 1129–1139. MR3010900

LIU, Y., ZHANG, H. H., PARK, C. and AHN, J. (2007). Support vector machines with adaptive L_q penalty. *Comput. Statist. Data Anal.* **51** 6380–6394. MR2408601

LOUW, N. and STEEL, S. J. (2006). Variable selection in kernel Fisher discriminant analysis by means of recursive feature elimination. *Comput. Statist. Data Anal.* **51** 2043–2055. MR2307560

MICCHELLI, C. A., XU, Y. and ZHANG, H. (2006). Universal kernels. *J. Mach. Learn. Res.* **7** 2651–2667. MR2274454

MUNDRA, P. and RAJAPAKSE, J. C. (2010). SVM-RFE with MRMR filter for gene selection. *IEEE Trans. Nanobiosci.* **9** 31–37.

RAKOTOMAMONJY, A. (2003). Variable selection using SVM-based criteria. *J. Mach. Learn. Res.* **3** 1357–1370. MR2020764

SCHIELE, B. and CROWLEY, J. L. (1996). Object recognition using multidimensional receptive field histograms. In *Computer Vision ECCV'*96 610–619. Springer, Berlin.

STEINWART, I. and CHRISTMANN, A. (2008). *Support Vector Machines*. Springer, New York. MR2450103

STEINWART, I. and SCOVEL, C. (2007). Fast rates for support vector machines using Gaussian kernels. *Ann. Statist.* **35** 575–607. MR2336860

SWAIN, M. J. and BALLARD, D. H. (1992). Indexing via color histograms. In *Active Perception and Robot Vision* 261–273. Springer, Berlin.

TANG, Y., ZHANG, Y. Q. and HUANG, Z. (2007). Development of two-stage SVM-RFE gene selection strategy for microarray expression data analysis. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **4** 365–381.

TSYBAKOV, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *Ann. Statist.* **32** 135–166. MR2051002

WANG, L., ZHU, J. and ZOU, H. (2006). The doubly regularized support vector machine. *Statist. Sinica* **16** 589–615. MR2267251

WESTON, J., ELISSEEFF, A., SCHÖLKOPF, B. and TIPPING, M. (2003). Use of the zero-norm with linear models and kernel methods. *J. Mach. Learn. Res.* **3** 1439–1461. MR1983013

ZHANG, H. H., AHN, J., LIN, X. and PARK, C. (2006a). Gene selection using support vector machines with non-convex penalty. *Bioinformatics* **22** 88–95.

ZHANG, X., LU, X., SHI, Q., XU, X., HON-CHIU, E. L., HARRIS, L. N., IGLEHART, J. D., MIRON, A., LIU, J. S. and WONG, W. H. (2006b). Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data. *BMC Bioinform.* **7** Art. ID 197.

ZHU, J., ROSSET, S., HASTIE, T. and TIBSHIRANI, R. (2003). 1-norm support vector machines. In *Neural Information Processing Systems* 16 49–56. MIT Press, Cambridge, MA.

S. DASGUPTA
M. R. KOSOROK
DEPARTMENT OF BIOSTATISTICS
UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL
CHAPEL HILL, NORTH CAROLINA 27599
USA
E-MAIL: sdg.roopkund@gmail.com
kosorok@unc.edu

Y. GOLDBERG
DEPARTMENT OF STATISTICS
UNIVERSITY OF HAIFA
MOUNT CARMEL, HAIFA 31905
ISRAEL
E-MAIL: ygoldberg@stat.haifa.ac.il