Master's thesis

Investigating a new approach to classification from statistical learning theory

Conditional probability estimation based on statistical invariants

Institute of Statistics
Ludwig-Maximilians-University Munich

submitted by

Felix Völpel

June 20, 2022

Supervisors:	Prof. Thomas Augustin Prof. Georg Schollmeyer

Declaration of Authorship

I hereby declare, that the thesis submitted is my own unaided work. All direct or indirect sources used are acknowledged as references.

This paper was not previously presented to another examination board and has not been published.

Munich, June 20, 2022

Felix Völpel

Abstract

The thesis presents a comprehensive study of the LUSI method introduced by Vapnik and Izmailov for the binary classification problem. Different than usual statistical approaches to classification, it results from a more analytical approach of approximating the conditional probability function. The method bases on an inverse-problem definition of the desired function and the appropriate application of Tikhonov's regularization principle. It is complemented by so called statistical invariants. Provided is a rework of related theoretical concepts. In a practical assessment the method is evaluated at the side of other models regarding the quality of function approximation and the accuracy of classification. While the basic approach could be verified to improve the quality of estimating the conditional probability function, the statistical invariants have not shown a systematic improvement, except for single cases. Regarding the classification accuracy the method did not establish any improvements compared to the common models such as empirical risk minimization.

Contents

1 Introduction

Various methods for binary classification, ranging from statistical estimation such as logistic regression to plenty methods being matter of machine learning, have been developed over the last decades. Most often these methods base on a statistical approach of explaining the binary observations conditioned on additional features by decoupling the stochasticity. Examples include maximum likelihood estimation or squared-loss minimization. From a stochastic point of view, the binary observation conditioned on additional variables is best characterized by its conditional probability distribution. Therefore, a new method has been proposed by Vapnik and Izmailov, which is constructed to explicitly approximate this function. The method bases on an inverse-problem definition of the conditional probability function, to derive an optimization problem on the basis of a given random sample. This is complemented by so called statistical invariants, which serve the incorporation of additional information about the desired function. The estimated conditional probability function is used to derive a decision rule to perform classification. In summary, the method is intrinsically different as it comes from a function approximation approach.

The purpose of this thesis is to provide a comprehensive investigation of this new approach to classification. This starts of with a profound analysis of the theoretical foundations, which also build the mathematical cornerstones of most models for empirical dependency estimation. The idea is to provide a summary of the main important conclusions of theoretical literature, working out a theoretical background in which among others the investigated method is perfectly embedded. In a second part, the new model is explicitly defined and the various notions behind it explained and further discussed. The last part deals with a practical evaluation. The authors presented a simulation based study in the original paper. The objective is to verify these results by reconstructing and extending their study. As will be pointed out, the analysis conducted in this thesis significantly augments the reliability of the provided statements. In the evaluation the method is compared with other models regarding two aspects. This is once the quality of approximating the conditional probability function and secondly the accuracy of the subsequent classification.

2 Theoretical foundations

The purpose of this chapter is to give an insight into selected aspects of ill-posed inverse problems, statistical learning theory and reproducing kernel Hilbert spaces. These concepts build the theoretical foundation of many statistical learning problems within which the later presented method can perfectly be reflected. Presented are the leading ideas and relevant relationships. Important proofs are outlined and references for detailed derivations given. In the appendix a summary of some of the main definitions of functional analysis is provided (app.A).

2.1 Ill-posed problems

In the following section the notion of ill-posed problems together with the most common solution framework by Tikhonov are introduced.

2.1.1 Definition and inverse problems

The central concept to define is the notion of a well-posed problem, which was first introduced by Jacques Hadamard. However, at first the term "problem" has to be clarified. Originally the concept was introduced in the context of differential equations. But the (mathematical) formalization of a problem is not necessarily restricted to this. The definition of a general problem introduced here is unique to this thesis and shall unify different definitions found in the literature. Let $(\mathcal{G}, \mathrm{d}_{\mathcal{G}})$ and $(\mathcal{Z}, \mathrm{d}_{\mathcal{Z}})$ be two metric (topological) spaces. $\mathcal{G}$ is called the data space and $\mathcal{Z}$ the solution space. A problem is identified with a subset $\mathcal{P}$ of $\mathcal{G} \times \mathcal{Z}$, which defines the set of admissible pairs, which implies the problem: Identify for a given g a z, such that (g, z) lies in $\mathcal{P}$. A stated problem is called well-posed, if the following three properties hold [1, 2]:

1. For each g in $\mathcal{G}$ there exists a solution z in $\mathcal{Z}$ such that (g, z) in $\mathcal{P}$.
2. For each g in $\mathcal{G}$ the corresponding solution z is unique, i.e. if (g, z_1), (g, z_2) in $\mathcal{P}$ then $z_1 = z_2$.
3. The function $L : \mathcal{G} \to \mathcal{Z}$, $g \to L(g)$ with $(g, L(g))$ in $\mathcal{P}$ is continuous.

A problem is called ill-posed, if one of the conditions is not fulfilled. It is called essentially ill-posed, if none of the conditions are fulfilled. However, the most critical aspect is usually the continuity. Note, that the function L is well-defined due to the first two properties. Well-posed means, that the problem admits a continuous solution function L, which is defined for all problem instances g of $\mathcal{G}$. Existence and uniqueness essentially ensure that the problem is (uniquely) solvable. The continuity is required to guarantee a certain stability of the solution $L(g)$ under small disturbances of g. The relevance of continuity will become obvious in the context of measurements.
For the computational problem of evaluating a function $A : \mathcal{D}_A \to \mathcal{Z}, \mathcal{D}_A \subseteq \mathcal{G}$ the three properties are satisfied, if $\mathcal{D}_A = \mathcal{G}$ and A is continuous (uniqueness is clear if A is a function). This kind of problem formulation is sometimes called direct problem (the function A is known). It holds $L = A$. Correspondingly there is a kind of problem, which is referred to as inverse problem. An inverse problem is defined implicitly by a function $A : \mathcal{Z} \to \mathcal{G}$, such that for a given instance g out of G, the task consists in determining the z such that $A(z) = g$. For the inverse problem, existence and uniqueness do ensure the existence of an inverse function A^{-1} such that $A(A^{-1}(g)) = g$. Then being well-posed equates to the direct problem associated to A^{-1} being well-posed. However, note that in general A being well-posed does not imply A^{-1} to be well posed (assuming its existence). This is because the inverse of a continuous function is not necessarily continuous. Inverse problems are often ill-posed and even essentially ill-posed. But inverse problems take a central role in applied sciences because they represent the typical approach of modelling. Assuming any kind of system, which transforms a causation z to an effect g. Mostly the challenge to any scientist is to get to know the causation, the present state of a system, but having only access to its effect g. However, it is much harder to relate from an effect to a specific, unique causation. Therefore, a

model is developed in the inverse direction, by formulating a law A, which relates a causation z to its effect $A(z)$. A direct problem could instead be interpreted as predicting the future state of a system, considering its present conditions. Often the spaces $(\mathcal{G}, \mathrm{d}_{\mathcal{G}})$ and $(\mathcal{Z}, \mathrm{d}_{\mathcal{Z}})$ are assumed to be normed vector spaces and the function A a linear, continuous operator. For the inverse problem a so-called linear operator equation does result. In [1] a list of examples of ill-posed direct and inverse linear operator problems is specified. The most prominent examples of ill-posed inverse problem are integral equations. Correspondingly for direct problems this is the linear differential operator. It being ill-posed shall be proven by a simple example:
Consider $A \colon \mathrm{C}^1([0,1]) \to \mathrm{C}([0,1])$, $f \to \frac{\mathrm{d}f}{\mathrm{d}x}$ and both spaces are equipped with the supremum's norm. Define the series $f_n(x) = x^n$, then $A(f_n) = n\, x^{n-1}$ for $n \in \mathbb{N}$. It holds: $\|A(f_n)\| = \sup_x(n\, x^{n-1}) = n \cdot 1 = n\, \|f_n(x)\| \implies$ there can't exist a constant N such that $\|A(f_n)\| \leqslant N\, \|f_n\|$ for all n. Thus, the operator A is unbounded and due to its linearity not continuous (app.A).
Special classes of operators are also compact operators and absolutely continuous operators, which don't have a continuous inverse operator and therefore imply ill-posed inverse problems [3, pp. 10, 20].

2.1.2 Solving ill-posed problems

The concept of well-posed problems has a very practical oriented motivation. Often the problem instance g, for which the solution should be found, is only accessible by measurements, hence $\mathcal{G}$ called the data space. However, the measurement process is typically incomplete, meaning that if g is a function itself, it can only be evaluated at finitely many points. Furthermore, the measurements are defective due to measurement errors leading to imprecise assessments. Consequently, instead of the real g, one obtains g_d, i.e. an element of $\mathcal{G}$ such that $\mathrm{d}_{\mathcal{G}}(g, g_d) \leqslant d$. This means, one is not even aware of the real instance g, which defines the actual, current task. Therefore, one can not really wish to find the desired $L(g)$, but only a good approximation to it. At this point the concept of a problem being well-posed becomes relevant. Intuitively the value $L(g_d)$ could be used as approximation of $L(g)$. If the problem is well-posed, then existence and uniqueness does ensure that

$L(g_d)$ is well-defined. Note, that existence is not granted due to possible measurement errors, which might cause g_d to fall out of $\mathcal{G}$ [1]. The continuity does justify this procedure as a reasonable strategy. Because as $d \to 0$ converges, the convergence of $L(g_d) \to L(g)$ is guaranteed. Note, that the problem being well-posed does not represent any constructive argument how to actually calculate $L(.)$. Especially in case of inverse problems it might be actually very difficult to determine a solution. However, if the problem is ill-posed then this convergence is not guaranteed, nor is the existence of $L(g_d)$. Thus, an approach has to be worked out how to deal with ill-posed problems. In the following, unless other stated the inverse problem formulation is considered, as these cover an essential part of ill-posed problems.

Obviously, the categorization of a problem whether well or ill-posed depends on the characterization of the spaces $\mathcal{G}$ and $\mathcal{Z}$. Thus, a general idea might be to restrict the spaces in a reasonable way. This leads to a slightly different concept, which is the conditionally well-posedness (or Tikhonov well-posed). An inverse problem is called conditionally well-posed with respect to a set $M \subseteq \mathcal{Z}$, if it is Hadamard well-posed regarding the spaces $(A(M), \mathrm{d}_{\mathcal{G}})$, $(M, \mathrm{d}_{\mathcal{Z}})$ [1, 3, 4]. Note, that continuity of L over $A(M)$ does not imply continuity in $(\mathcal{G}, \mathrm{d}_{\mathcal{G}})$. The set M is also called the set of correctness. Such a restriction represents having additional information about the problem state or making any assumptions, which lead to the conclusion that the solution z is element of a subset $M \subseteq \mathcal{Z}$. If M is a compact set and A is a continuous function (such as a bounded linear operator), then the problem is conditionally well-posed with respect to M. This results by the following relation: If $f : M \to f(M)$ is a bijective, continuous function over a compact set M, then $f^{-1} : f(M) \to M$ is continuous too. This is easily proven. Because f is continuous and M is compact any closed subset of M is a compact set and is therefore mapped to a compact, thus closed subset of $f(M)$. Hence, f as inverse of f^{-1} maps any closed set of M to a closed set in $f(M)$. However, defining a proper set M is usually quite difficult and restricting $\mathcal{G}$ to $(A(M), \mathrm{d}_{\mathcal{G}})$ might not be a good choice [5]. Firstly, verifying whether the unknown solution z is indeed lying in M is very hard (unless explicitly stated) [1]. Especially the calculation

[1] The problem definition is regarding $(\mathcal{G}, \mathrm{d}_{\mathcal{G}})$. But due to the measurements' error one actually operates in $\tilde{\mathcal{G}} \supseteq \mathcal{G}$. But for this space the existence might not be given anymore.

of $A^{-1}(g_d)$ does not reveal anything about the real $z = A^{-1}(g)$. Secondly, even if z is in M equivalently g being in $A(M)$, because of the measurement process it is not guaranteed that the observed g_d is element of $A(M)$. This is especially sensitive, as all approximations $g_d \to g$ must lie in $A(M)$, in order to benefit from the conditional continuity, i.e. the convergence of $A^{-1}(g_d)$ to z. Nevertheless, the notion of conditionally well-posedness shows, that in order to approach an ill-posed problem either more information or further assumptions are required. The construction and isolation of compact sets of admissible solutions is in fact the basis of many methods, to approximate ill-posed problems [6]. Some of them are the methods by Ivanov, Lavrentiev and Morozov. These rely mainly on deriving a so called "quasi-solution" and restricting the solution space. More information can be found in [7].

A widely applied method is Andrei Tikhonov's regularization principle. The method is even applicable to essentially ill-posed problems and has various advantages, which will be presented in the following. Tikhonov's solution does not rely on restricting the set of admissible solutions. Instead, it bases on the assumption, that information about the measurement error d between g_d and g is available. The essential concept is the notion of the "regularization operator" [2], which formalizes what a "good" approximation algorithm is. For any ill-posed problem, a parametric regularization operator $R : \mathcal{G} \times \mathbb{R}^+ \to \mathcal{Z}$ is defined by the following properties [6, 1] [3]. Let g and $z = L(g)$ be arbitrary yet fixed instances:

1. The functions $R(., \gamma)$ are continuous in g.

2. $R(g, \gamma) \to z$ as $\gamma \downarrow 0$

3. There exists an increasing function $\gamma(d)$, with $\gamma(d) \to 0$ as $d \to 0$, such that for any $\epsilon > 0$, there is a $d_0(\epsilon) > 0$ and $\mathrm{d}_{\mathcal{G}}(g_d, g) < d_0(\epsilon) \implies \mathrm{d}_{\mathcal{Z}}(R(g_d, \gamma(d_0)), z) < \epsilon$ holds for all such g_d.

A problem is called regularizable if there exists such a regularization operator. As turns out, many practical problems are regularizable. In fact, there are also

[2] In some literature this is also called the "regularization algorithm" or "-functional".

[3] In [5] this is defined by $\lim\limits_{\gamma \to 0} \sup\limits_{g_d,\, d_{\mathcal{G}}(g_d, g) < \gamma} d_{\mathcal{Z}}(R(g_d, \gamma), L(g)) = 0$, the parametrization is d itself.

naturally occurring regularization operators. One example is the difference quotient, as regularization operator for the ill-posed differential operator. Note, that for a well-posed problem L itself is a regularization operator. For an ill-posed, regularizable problem the functions R_γ can be interpreted as continuous approximations of L, which converge pointwise to L. However, to ensure the convergence of $R(g_{d_i}, \gamma)$ for an arbitrary sequence of approximations $g_{d_i} \rightarrow g$ towards $z = L(g)$, the γ parameter has to be chosen appropriately in dependence of d. According to Tikhonov $R(g_d, \gamma(d))$ is then a reasonable approximation for z with d being the data error. Therefore, the problem of solving ill-posed problems reduces to find a regularization operator and a dependency relation $\gamma(d)$. The role of $\gamma(d)$ becomes more clear, when one tries to estimate the error between z and $z_{d,\gamma} = R(g_d, \gamma)$. To do so, consider the following inequality with $z = L(g)$: $\mathrm{d}_{\mathcal{Z}}(z,\, z_{d,\gamma}) \leqslant \mathrm{d}_{\mathcal{Z}}(z,\, R(g, \gamma)) + \mathrm{d}_{\mathcal{Z}}(R(g, \gamma),\, z_{d,\gamma})$. For simplicity, let $\mathcal{G}$ and $\mathcal{Z}$ be Banach spaces and R_γ a linear operator for any γ. Then, $\mathrm{d}_{\mathcal{Z}}(z,\, R(g, \gamma)) = \|z - R(g, \gamma)\| =: \rho(z, \gamma)$. Furthermore, $\mathrm{d}_{\mathcal{Z}}(R(g, \gamma),\, z_{d,\gamma}) = \|R(g, \gamma) - R(g_d, \gamma)\| = \|R(g_d - g, \gamma)\| \leqslant \|g_d - g\|\,\|R_\gamma\| \leqslant d\,\|R_\gamma\|$. It follows, $\|z_{d,\gamma} - z\| \leqslant \rho(z, \gamma) + \|R_\gamma\|\, d$ [1, p. 349]. This inequality is representative for a well-known trade-off. Obviously, the error between the approximation $z_{d,\gamma}$ and z should be as small as possible. $R(g, \gamma)$ is the approximation in the best case scenario of knowing the real g. The trade-off results from seeking an approximation algorithm, whose output does not vary to much, if a data error d is present compared to its value at g (continuity). But which is still close to the real solution z, if the error actually decreases/vanishes (closeness to L). Due to the definition of R it follows $\rho(z, \gamma) \rightarrow 0$ for $\gamma \rightarrow 0$. However, because the R_γ are continuous and converge pointwise to L, it follows that the R_γ can not converge uniformly. Otherwise, L would be continuous as well meaning that the problem would not be ill-posed. This implies, that $\|R_\gamma\| \rightarrow \infty$ for $\gamma \rightarrow 0$. Thus, one has to balance between both errors. The parameter γ has to be chosen in dependence of d and steers to what extent one relies on L.

An important case is that of $\mathcal{G}$ and $\mathcal{Z}$ being Hilbert spaces. Then, Tikhonov's theory proposes the construction of a regularization operator in the following way: $R(g, \gamma) := \underset{z}{\operatorname{argmin}} \|A(z) - g\|_{\mathcal{G}}^2 + \gamma \|z\|_{\mathcal{Z}}^2$. If A is a linear, compact operator one can prove that the minimizer exists, is unique and depends continuously on g.

Thus, it is in fact a regularization operator. Regarding the choice of the parameter dependency $\gamma(d)$, the following is proven: If $\gamma(d)$ satisfies 1) $\lim\limits_{d\to 0}\gamma(d) = 0$ and 2) $\lim\limits_{d\to 0}\frac{d^2}{\gamma(d)} = 0$, then for any sequence $g_{d_k} \to g$ the minimizer converges $R(g_{d_k}, \gamma(d_k)) \to A^+g$, with A^+ being the generalized inverse operator (Moore-Penrose Operator) [1, 8]. This statement admits several insights. Firstly, the requirements on $\gamma(d)$ to ensure convergence are comparatively simple, which constitutes a first advantage of the method. Secondly, the regularization operator does approximate the operator A^+. This generalized inverse operator is well-defined for all linear operators A. The operator A^+ is discontinuous if and only if the range of A denoted by $\mathrm{rng}(A)$ is not closed. This is for example given, if A is compact and the dimension of $\mathrm{rng}(A)$ is infinite, which frequently occurs. Hence, the problem of evaluating A^+ is usually ill-posed. The domain of A^+ is defined by $D(A^+) = \mathrm{rng}(A) + \mathrm{rng}(A)^\perp$. This does not necessarily equal $\mathcal{G}$. The sets are different if $\mathrm{rng}(A)$ is not closed [8, 9]. It can be shown, that for all $\mathcal{Z}\backslash D(A^+)$ the sequence of $R(g_d, \gamma(d))$ does not converge. If A^+g does exist, it is the so called minimum norm quasi solution, i.e. the $\tilde{z}$ with the smallest norm that minimizes $\|Az - g\|^2_{\mathcal{Z}}$. Hence, if the problem does not admit the existence of a solution a reasonable alternative is defined. This represents another characteristic of the Tikhonov regularization, as it is applicable to essentially ill-posed problems. Note, that if there is a z in $\mathcal{Z}$, such that $Az = g$ holds, then the minimum norm solution is an exact solution. Furthermore, the operator A^+ can be expressed analytically, so that the optimization problem has a closed form solution: $R(g_d, \gamma) = (A^*A + \gamma E)^{-1}A^*g_d$, at which A^* is the adjoint operator of A (definition of adjoint in app.A).

Some adaptions of the regularization operator have to be made, when encountering other assumptions on A, $\mathcal{G}$ and $\mathcal{Z}$. If A is not linear, then the minimizer might not be unique anymore. In this case only the existence of a convergent subsequence can be proven [10]. Furthermore, in case of general metric spaces, the norm is replaced by the corresponding metric, while the term around z becomes a general so called stabilizing functional $O(z)$. More information can be found in [3]. A relevant generalization worth to mention is the following. It might not be possible to even evaluate the real A, but only an approximate A_h, where $\|A - A_h\| \leqslant h$ with respect to some operator norm. Again, the same regularization operator is applicable by substituting A with its approximation A_h. To guarantee the convergence of the sequences $R(g_d, \gamma, h)$, as $h, d \to 0$ some further requirements on $\gamma(d)$ and $h(d)$ are required. Detailed information can be found in [1, 11].

2.2 Statistical learning theory

The following section deals with the backgrounds of the statistical learning theory (SL-theory) mainly shaped by Vladimir Vapnik and Alexey Chervonenkis. The theory is a mathematical framework, which analyses the conditions under which the estimation of dependencies based on empirical data succeeds. Specifically, it investigates the fundamental concept of empirical risk minimization based on which the extension of structural risk minimization is suggested. For a detailed insight the following literature can be consulted, based on which the here provided information was obtained [3, 4, 12].

2.2.1 Risk functional and problems of dependency estimation

In many use cases of natural sciences or engineering the mathematical abstraction of a problem results in the optimization of a functional $I : \mathcal{F} \to \mathbb{R}$ over a set of functions $\mathcal{F}$. The elements of $\mathcal{F}$ are also called hypotheses. The functional is problem specific and represents the quality of a hypothesis $f \in \mathcal{F}$ according to the requirements of the problem. It is often encountered, that the functional I is in the form of an integral property, i.e. $I(f) = \int_{D_v} Q(v, f)\, \mathrm{d}v$. $Q(v, f)$ measures the quality of the function f at the selective context defined by v. Usually $v \in D_v$ is a vector of $\mathbb{R}^{\mathrm{d}}$. The SL-theory concerns problems where the vector v is subject to some randomness or uncertainty. This means, that the quality $Q(v, f)$ is weighted according to a certain probability distribution $\mathrm{P}(v)$ with respect to a proper probability space over D_v. The functional I results to be of the form $I(f) = \int_{D_v} Q(v, f)\, \mathrm{dP}(v)$. Under these assumptions $Q(., f)$ is a random variable and $I(f)$ its expectation. Q is called the loss function and I the risk functional, as it represents the cumulative quality of the function f. The vector v is usually decomposed into two parts $v = (y, x) \in \mathcal{Y} \times \mathcal{X}$. The functions f are defined over $x \in \mathcal{X}$ and are determined by some kind of parametrization, i.e. $f_\alpha(x) \in \mathcal{F}$, $\alpha \in \mathcal{A}$.

The parametrization is allowed to be very general and does not mean any restriction in the generality of the problem. But the identification of a function f equates now to determining the index α, which is why the notation changes to $Q(v, \alpha)$ and $I(\alpha)$ respectively. Based on this framework three types of problems are differentiated, which often occur in statistical use cases according to [3, ch. 1]. As will be shown these problems are indeed special instances of the problem of minimizing a specific risk functional. Note, that the definition of the functional I is determined by four components, i.e. the domain $\mathcal{Y} \times \mathcal{X}$, the function space $\mathcal{F}$, the function Q and the distribution $\mathrm{P}(v)$.

Pattern recognition problem

The so-called pattern recognition problem is the formalization of the well-known task of binary classification. The vector x is the finite dimensional representation of some kind of object and usually element of $\mathcal{X} \subseteq \mathbb{R}^d$. The variable y takes on the values either zero or one and encodes the affiliation to one of two concepts (classes) of the respective object. The conditional probability distribution $\mathrm{P}(y \mid x)$ models the uncertainty in the classification of the object x. This might occur either due to randomly false assignments or due to the fact, that x is an incomplete representation of the actual object. In a more abstract sense $\mathrm{P}(v)$ might be interpreted in the following way. According to $\mathrm{P}(x)$ an object x occurs. Then an arbiter decides based on the conditional distribution $\mathrm{P}(y \mid x)$ to what class this object should be assigned to. Since y is of discrete nature the functions of $\mathcal{F}$ should be indicator functions mapping into $\{0, 1\}$. They represent a decision rule of how the vector x is assigned to one of the classes. The objective of the pattern recognition is to realize successful classification. According to [3] the quality of a decision rule is best defined in terms of the probability of incorrect classification regarding the distribution $\mathrm{P}(v)$. This probability is expressed by: $\int_{\mathcal{Y} \times \mathcal{X}} \mathbb{1}_{\{y \neq f(x)\}} \,\mathrm{dP}(v)$. By defining $Q(v, \alpha) := (y - f_\alpha(x))^2 = \mathbb{1}_{\{y \neq f(x)\}}$ the integral can be written as $\int_{D_v} Q(v, \alpha) \,\mathrm{dP}(v)$. Hence, the probability of false classification of a decision rule f_α is a risk functional $I(\alpha)$. It is easily proven that among the sets of all possible decision rules the function $f^*(x) = \underset{y}{\operatorname{argmax}} \mathrm{P}(y \mid x)$ leads to the smallest error probability. The corresponding risk value is called the Bayes error.

Interpreting results of direct experiments / Regression estimation
Object of interest is a specific function f^*, which arises problem dependent. In the case of regression the motivation is to extract a summary of a conditional probability function $\mathrm{P}(y \mid x)$ in terms of the conditional expectation, i.e. $f^*(x) := \mathbb{E}(Y \mid x)$. In a more general context, there already exists some lawlike functional relation f^* which is to be determined. The evaluation of the function at a point x is called an experiment. However, this evaluation is assumed to be imprecise which is why the output of the experiment becomes a random variable $Y \mid x$ and is described by a probability distribution $\mathrm{P}(y \mid x)$. On this distribution three assumptions are made: 1) $\mathbb{E}(Y \mid x) = f^*(x)$, 2) $\mathbb{V}ar(Y \mid x) < \infty$, 3) $Y \mid x_i$ and $Y \mid x_j$ are independent for $x_i \neq x_j$. The vector (y, x) is element of $\mathbb{R}^{d+1}$, implying that y is real-valued too. The goal is to retrieve the unknown function on the basis of the experiments. Accordingly, $\mathcal{F}$ is a set of real-valued functions. While in the previous problem the risk optimization arose naturally, it has to be justified in this case. A well-known choice of Q is again the quadratic loss: $Q(v, \alpha) := (y - f_\alpha(x))^2$. To see that the task of function estimation can indeed be reduced to the problem of risk minimization, the minimizer of the corresponding risk $I(\alpha)$ is analysed. One confirms, that the minimization of $I(\alpha)$ is obtained at the function in $\mathcal{F}$, which has the smallest distance to f^* with respect to $L^2_{\mathrm{P}(x)}$. Hence, by calculating the minimizer the desired function is reasonably approximated dependent on the choice of $\mathcal{F}$.

Although the settings are very similar, the objectives of both introduced problems are intrinsically different. In the pattern recognition problem the task was to optimize the classification quality expressed in terms of the probability of false classification, which is a risk functional itself. But there is no "true" decision rule [4]. In the context of evaluating direct experiments the task is to determine a specific function, which occurs to be the minimizer of an appropriate risk functional. Hence, the risk functional takes on different roles either as natural measure of quality or as tool to implicitly define the desired function in the context of risk minimization.

[4] Of course one could imagine a lawlike assignment procedure, which is supposed to be identified. The other way around one could define the regression setting as prediction task. Therefore, the distinction is not that much along the concrete problems but rather along the two different purposes of the risk functional.

Interpreting results of indirect experiments

The last dependence structure describes another case, where the objective lies in determining a specific function. But instead of having access to measurements of $f^*(x)$, one can only access a transformed version of it, i.e. $A(f^*) = F^*$. Therefore, the assumptions made on the measurements Y are changed to $\mathbb{E}(Y \,|\, x) = F^*(x)$. The choice of Q remains the same. The described scenario is this of an inverse problem, as introduced in the previous chapter. Obviously of special interest are ill-posed problems, such that the calculation of the inverse operator A^{-1} is not applicable.

So far three instances of the framework of risk minimization were presented. Minimizing the respective risk functionals is the objective but not the actual challenge the SL-theory has to deal with. This is given by the circumstance, that the distribution $\mathrm{P}(v)$ is not known, especially the conditional distribution $\mathrm{P}(y \,|\, x)$ [5]. The only available information about $\mathrm{P}(v)$ is a random sample of l i.i.d. observations denoted by $v^{(l)}$. Consequently, the risk functional can only be approximated. Hence, it cannot be demanded to really minimize the functional I based on incomplete information about it. Thus, the resulting statistical problem is less an optimization problem, but rather the search for an estimator (algorithm) $\hat{\alpha} : D_v^l \to \mathcal{A}$, which guarantees a sufficient closeness of $I(\hat{\alpha}(v^{(l)}))$ to the minimal value $I^*_{\mathcal{F}} := \min\limits_{f_\alpha \in \mathcal{F}} I(\alpha)$. This shifts the focus from comparatively simple optimization to the construction of an appropriate estimator. The quality of such an estimator is assessed based on the distribution of $I(\hat{\alpha}) - I^*_{\mathcal{F}} > 0$, which becomes a random variable. The SL-theory is especially devoted to making statements about the quantiles (or some upper bound of it) $\varkappa(\eta, l)$ of $I(\hat{\alpha}) - I^*_{\mathcal{F}}$ dependent on the confidence level η and the sample size l. An estimator with a smallest upper bound $\varkappa(\eta, l)$ for a given probability η and sample size l would then be optimal. Because $I(\hat{\alpha})$ remains unknown, an estimator should allow the calculation of a meaningful confidence interval. As will be seen, this requirement falls together with the derivation of $\varkappa(\eta, l)$. Of special interest is the convergence behaviour of such estimators.

[5] For the construction of algorithms and their investigation the assumptions about $\mathrm{P}(x)$ are not of such decisive relevance. According to [3] this makes the differentiation between "open" and "closed" world, at which the distribution $\mathrm{P}(x)$ is either known or unknown.

The estimator's risk $I(\hat{\alpha})$ should convergence in probability to the minimal value $I^*_{\mathcal{F}}$ in order to guarantee a consistent improvement for an increasing amount of information. Clearly the minimum $I^*_{\mathcal{F}}$ depends on the chosen function space. Thus, for different spaces a converging estimator converges to different minima. Therefore, a different perspective is considered. Let $\tilde{\mathcal{F}}$ be an overall embedding function space of admissible functions. The corresponding minimum of the risk functional is the globally possible minimum $I^*_{\tilde{\mathcal{F}}}$. Hence, regarding any approximation the relevant quantity is the difference $I(\hat{\alpha}) - I^*_{\tilde{\mathcal{F}}}$. However, for the actual construction of an estimator only a subspace $\mathcal{F} \subset \tilde{\mathcal{F}}$ is considered. Therefore, the difference can be decomposed in the following way: $I(\hat{\alpha}) - I^*_{\tilde{\mathcal{F}}} = (I(\hat{\alpha}) - I^*_{\mathcal{F}}) + (I^*_{\mathcal{F}} - I^*_{\tilde{\mathcal{F}}})$. This representation is sometimes referred to as "generalization - approximation trade-off" [13]. The first term quantifies the closeness to the achievable minimal risk value for the estimator operating in $\mathcal{F}$. While the second term describes the discrepancy between the global minimum and the local minimum. As will be seen, this results in a trade-off, as more complex spaces do enable a smaller approximation error but cause a worse convergence behaviour on the other side. This relation already indicates that the bound $\varkappa(.)$ must depend on some kind of characteristic of the used function space $\mathcal{F}$. An estimator which yields a comparatively small bound $\varkappa$ on $I(\hat{\alpha}) - I^*_{\mathcal{F}}$ is called to generalize well. However, this only means, that the associated risk value is close to the possible (local) minimum, but not necessarily small.

Because $f_{\hat{\alpha}(v^{(l)})}$ as well as $I(\hat{\alpha}(v^{(l)}))$ can only be approximations of the "true" values, it is required to reevaluate the role of the risk functional in general. Whether the construction of I is only an auxiliary concept or is it the intrinsic indicator of success? Thus, is it about the actual minimizer or just the minimization? For tasks of prediction such as the pattern recognition problem, the risk functional expresses the immediate quantity of interest, as described before. Therefore, two decision rules are treated as similar if their risk functional's value is similar too. However, this is different if the focus lies on the real functional relation, as originally stated for the problems of direct and indirect experiments. The derived risk functional served the purpose of implicitly defining the desired function as its minimizer. But the similarity of functions is rather be defined by the common metrics such as the L^2 metric or supremum's norm. However, the convergence of risk values do not necessarily imply the convergence of the respective functions. In the case of regression estimation, it holds per definition that similarity in the defined risk between two functions implies closeness with respect to the $L^2_{\mathrm{P}(x)}$ metric. However, regarding the supremum's norm this is not anymore guaranteed. Especially in case of ill-posed inverse problems, the convergence to none of these metrics can be assured on the basis of similar risks, as described in the previous chapter. Therefore, one should be aware of the adjusted objective due to the statistical challenge. This is for a given sample to obtain, with a certain guarantee in probability, a value close to the risk functional's minimum. Consequently the similarity of different functions of $\mathcal{F}$ is measured by means of the respective risk functional.

2.2.2 Uniform convergence

A generic idea for constructing an estimator $\hat{\alpha}$, lies in constructing an estimate $\hat{I}$ of the risk functional I on the basis of the given sample, whose minimization defines the estimator $\hat{\alpha} = \underset{\alpha}{\operatorname{argmin}} \hat{I}(\alpha)$. According to the previous section, the value of interest is $I(\hat{\alpha})$, i.e. the risk value of the approximated minimizer. As stated, the requirements of a reasonable estimator are the convergence of $I(\hat{\alpha})$ to the minimal risk value and the knowledge about an upper bound of it with a guaranteeing probability. Both questions are approached by deducing a sufficient condition for convergence. For this purpose, an upper bound on the value $|I(\hat{\alpha}) - I^*_{\mathcal{F}}|$ is derived. Let α^* be the minimizer of the risk functional among $\mathcal{F}$, i.e. $I(\alpha^*) = I^*_{\mathcal{F}}$, then it holds:

$$\begin{aligned} |I(\hat{\alpha}) - I^*_{\mathcal{F}}| &\leqslant |I(\hat{\alpha}) - \hat{I}(\hat{\alpha}) + \hat{I}(\alpha^*) - I(\alpha^*)| && (1) \\ &\leqslant |I(\hat{\alpha}) - \hat{I}(\hat{\alpha})| + |\hat{I}(\alpha^*) - I(\alpha^*)| && (2) \\ &\leqslant 2\,\tau \text{ with } \tau := \sup_{\alpha} |I(\alpha) - \hat{I}(\alpha)| < \infty && (3) \end{aligned} \tag{2.1}$$

Step (1) is valid because $\hat{a}$ is the minimizer of the estimated functional $\hat{I}$, thus a positive term is added. (2) follows by definition of any metric. The assumption which has to be made is mirrored in step (3), i.e. the existence of a finite supremum between the true risk functional and its estimate. The supremum is a measure of the quality of the approximation of I, as it is its worst-case deviation. The assumption of the existence of a supremum is usually propagated back to assumptions about the loss function Q (and the function space $\mathcal{F}$). By the inequality follows, if the supremum converges towards zero, then the risk of the estimate $\hat{\alpha}$ becomes arbitrarily close to this of the real minimium. The convergence is formalized by convergence in probability for random variables, i.e. $\forall\, t \in \mathbb{R}^+ : \mathrm{P}(\sup_\alpha |I(\alpha) - \hat{I}(\alpha)| > t\,) \to 0$ as $l \to \infty$ [6]. This property of any estimator $\hat{I}$ is called uniform convergence [3, ch. 2.6]. In order to construct a confidence interval, it is sufficient to (non-trivially) bound this probability, i.e. $\mathrm{P}(\,\tau > \varkappa\,) < 1 - \eta(\varkappa, l)$ (with $\eta(\varkappa, l) \to 1$ as $l \to \infty$ for any $\varkappa$). By deriving $\eta = 1 - \eta(\varkappa, l) \iff \varkappa = \varkappa(\eta, l)$ this is used to construct the following interval. It holds $\tau < \varkappa \implies |I(\hat{\alpha}) - \hat{I}(\hat{\alpha})| < \varkappa$, hence for the sample size l the following interval is valid with the probability of at least η: $\hat{I}(\hat{\alpha}) - \varkappa(\eta, l) < I(\hat{\alpha}) < \hat{I}(\hat{\alpha}) + \varkappa(\eta, l)$. This bound might be coarse, as it is valid for arbitrary α at the same time. In order to derive a reasonable bound on this probability some kind of assumptions have to be made specifically about the distribution of $Q(., \alpha)$. The minimum amount of information is in terms of a uniformly bounded variance (or variation coefficient) in α or the uniform boundedness of Q itself [7]. The last falls together with the assumption of the finiteness of τ. These are rather weak assumptions comparing with any parametric distributions. In fact, in [3, ch. 2] it is shown, that various distributions $\mathrm{P}(v)$ can be grouped behind the restriction of a uniformly bounded variance of $Q(., \alpha)$. Since these are the only required assumptions the validity of the SL-theory is mostly independent of the concrete shape of Q.

Uniform convergence is a sufficient condition for an estimator of I to imply a reasonable estimator $\hat{\alpha}$. But there are multiple ways, how to estimate the risk functional. The risk value of a function is an expectation value with respect to a probability distribution $\mathrm{P}(v)$, which can be written in terms of the corresponding density function, presupposing its existence. Thus, one approach could be estimating the density function and its subsequent substitution in the risk functional's definition. This idea bases on two observations. First, the continuity of the integral opera-

[6] Although in the notations it is omitted, but the randomness of $\hat{I}(\alpha)$ occurs, as the estimate $\hat{I}$ depends on the random sample $v^{(l)}$.

[7] This is because intervals can then be derived based on Chebychev's or Hoeffding's inequalities.

tor, which ensures that the closer the estimate of the density function, the closer the risk estimate to the real risk value. Second, the well-known Glivenko-Cantelli theorem, which asserts a sufficiently good estimator of the distribution function and might therefore indicate the existence of a reasonable estimator of the density function too. However as turns out, estimating the density function without any prior information is a rather complex (ill-posed) problem [3, ch. 2.5]. Thus, there is no real gain in reducing the task of estimating the risk functional to the problem of estimating the density function. The method which is finally applied is the so called functional of empirical risk, i.e. $\hat{I}(\alpha) = I_{emp}(\alpha) = \frac{1}{l}\sum_{i=1}^{l} Q(v_i, \alpha)$. The risk value for a function f_α is estimated by the associate average value of the loss function Q over the given sample $v^{(l)}$. The corresponding minimizer is denoted by $\alpha_{emp} := \underset{\alpha}{\operatorname{argmin}}\, I_{emp}(\alpha)$. The SL-theory deals with the analysis of the empirical risk minimizer regarding conditions for uniform convergence for each of the previously presented problems of dependency estimation.

2.2.3 Sufficient conditions for uniform convergence

Pattern recognition problem

The foundation of the SL-theory lies in the analysis of the pattern recognition problem. At first one notices, that due to the discreteness of y and f_α the loss function $Q(v, \alpha)$ is bounded, as it takes on the values zero or one. Thus, the mentioned required assumptions are canonically satisfied. A simple situation is when the function space $\mathcal{F}$ is finite $N := |\mathcal{F}|$. Then the required probability regarding the supremum $\tau_{emp} := \sup_{\alpha} |I(\alpha) - I_{emp}(\alpha)|$ can easily be bounded. It follows:

$$\begin{aligned}
\mathrm{P}(\tau_{emp} > t) &= \mathrm{P}\Big(\bigcup_{n=1}^{N} \{v^{(l)} \,|\, |I_{emp}(\alpha_i) - I(\alpha_i)| > t\}\Big) && (1)\\
&\leqslant \sum_{i=1}^{N} \mathrm{P}(\,|I_{emp}(\alpha_i) - I(\alpha_i)| > t\,) && (2)\\
&\leqslant N\, 2\, exp(-2\, l\, t^2) && (3) \qquad (2.2)
\end{aligned}$$

It holds $N\, 2\, exp(-2\, l\, t^2) \to 0$ for $l \to \infty$, thus the uniform convergence is always satisfied in case of finitely many functions. The first equality follows by definition of the supremum. Step (2) results due to the sigma-sub-additivity of the probability measure. (3) is an immediate consequence of Hoeffding's inequality, which is applicable because of the relation $\mathbb{E}(I_{emp}(\alpha_i)) = I(\alpha_i)$ and the boundedness of I_{emp} between zero and one. The Hoeffding bound is independent of the α_i, which is why it can be summed over $\mathcal{F}$. Note, that in principle different bounds on the probabilities at (2) are possible [8]. However, the Hoeffding inequality requires only the boundedness, which is intrinsically satisfied for the pattern recognition problem. With this bound a confidence interval can be derived, as explained in the previous section. This simple bound already reveals, that the bigger N, the slower is the convergence of the probability. This indicates, that in case of infinitely sized sets $\mathcal{F}$ the uniform convergence must be tied to some further conditions about the function space.

To make statements about the general case of infinitely sized sets $\mathcal{F}$ a corresponding but more general problem is analysed. This is the uniform convergence of empirical frequencies of events to their probabilities. Let $\mathcal{E}$ be a set of arbitrarily many events with respect to a probability space $(D_v, \Sigma, \mathrm{P}(v))$. Let $v^{(l)} = (v_1, ..., v_l)$ denote the l dimensional random vector regarding the product measure associated to $\mathrm{P}(v)$, representing the set of possible samples of size l. Then $\nu_l(E) := \frac{1}{l}\sum_{i=1}^{l} \mathbb{1}_E(v_i)$ is the empirical average of the occurrences of event $E \in \mathcal{E}$ in a given sample. ν_l is said to uniformly converge if $\sup\limits_{E\in\mathcal{E}} |\nu_l(E) - \mathrm{P}(E)| \xrightarrow{\mathrm{P}} 0$ holds. Clearly I_{emp} is a special case of ν_l with $\mathcal{E}_Q := \{\{v \in \{0,1\} \times \mathbb{R}^d \,|\, Q(v, \alpha) = (y - f_\alpha(x))^2 = 1\} \,|\, f_\alpha \in \mathcal{F}\}$. $\mathcal{E}_Q$ is the set of sets of differently classified observations than observed applying any decision rule of $\mathcal{F}$. Central element of the study of uniform convergence is the so-called "growth function" of the set $\mathcal{E}$, which is denoted by $m_{\mathcal{E}}(l)$. By definition, an event E is a subset of a ground set. Thus, given a set of observations $\{v_1, ..., v_l\}$ each element is either element of E or not. Hence, the event E splits up a certain subset, which is $E \cap \{v_1, ..., v_l\}$. An example is given in table 2.1, where $D_v = \mathbb{N}$. Thus, the set $\mathcal{E}$ implies a number

[8] For example each probability could be bounded according to the Chebyshev inequality, which requires knowledge about the variance $\mathbb{V}\mathrm{ar}(Q(v, \alpha))$. These probabilities could then be bounded, if one assumed a known bound on the variances $\sup_\alpha \mathbb{V}\mathrm{ar}(Q(v, \alpha))$. Because Q is a Bernoulli variable the variance is bounded by 1/4.

	$v_1 = 1$	$v_2 = 2$	$v_3 = 3$	
$E_1 = \{1, 2, 4\}$	1	1	0	$\{v_1, v_2\}$
$E_2 = \{1, 2, 5, 6\}$	1	1	0	$\{v_1, v_2\}$
$E_3 = \{2, 3, 4\}$	0	1	1	$\{v_2, v_3\}$
...	...	...	...	...

Table 2.1: Three different events implying subsets on a given sample.

of subsets of $\{v_1, ..., v_l\}$ which can be differentiated by its events. This concept allows to define dependent on the given sample an equivalence relation $\sim_{v^{(l)}}$ over $\mathcal{E}$. Accordingly, two events are equivalent, if they imply the same subsets. Let $\kappa_{\mathcal{E}}(v^{(l)}) = |\{E \cap v^{(l)} \mid E \in \mathcal{E}\}|$ denote the number of these equivalence classes for the given set of l observations. $\kappa_{\mathcal{E}}(v^{(l)})$ is trivially bounded by 2^l. In case $\mathcal{E}$ is finite, the number of distinct subsets is bounded by $\min(2^l, |\mathcal{E}|)$. Note, that the set of equivalence classes alters depending on the concrete sample $v^{(l)}$. The growth function is defined as $m_{\mathcal{E}}(l) = \max\limits_{v^{(l)}} \kappa_{\mathcal{E}}(v^{(l)})$, i.e. the maximum number of differentiable subsets achievable over all possible samples of size l. The growth function is monotonically increasing. Further analysis reveals, that $m_{\mathcal{E}}(l)$ equals 2^l and is for $l > h_{\mathcal{E}}$ bounded from above by $l^{h_{\mathcal{E}}+1} + 1$, at which $h_{\mathcal{E}} = \max\{l \in \mathbb{N} \mid m_{\mathcal{E}}(l) = 2^l\}$ [12]. It is decisive that the existence of $h_{\mathcal{E}}$ is not guaranteed for each set $\mathcal{E}$. The value $h_{\mathcal{E}}$ is called the capacity (VC-dimension) of the set $\mathcal{E}$ and is set to infinity, if it does not exist. It can be interpreted as the effective size in the sense of diversity of $\mathcal{E}$ regarding its "power" to separate elements of a sample. The uniform convergence is investigated by means of an additional lemma. The lemma does replace the required supremum over $\mathcal{E}$ by a supremum over a finite (yet random) set of random variables. More precisely, according to [12] for $l > \frac{2}{t^2}$ it holds: $\mathrm{P}(\sup\limits_{E} |\nu_l(E) - \mathrm{P}(E)| > t) \leqslant 2\,\mathrm{P}(\sup\limits_{E} |\nu_l(E) - \tilde{\nu}_l(E)| > \frac{t}{2})$. $\tilde{\nu}_l$ is the analogue function but for a second, independent sample of size l. The variable $|\nu_l(E) - \tilde{\nu}_l(E)|$ maps into the same range of values $\{\frac{1}{l}, \frac{2}{l}, ..., 1\}$ for all events E and samples $v^{(2l)}$. Furthermore, for a given sample of size $2l$ two events E_1, E_2 which fall into the same equivalence class do evaluate to the same value. Thus, the supremum can be related to the finite set of equivalence classes: $\sup\limits_{E} |\nu_l(E) - \tilde{\nu}_l(E)| = \sup\limits_{\mathcal{E} \backslash \sim_{v^{(2l)}}} |\nu_l(E) - \tilde{\nu}_l(E)|$. Taking into account, that the number of equivalence classes is bounded by the growth function further elaborating the inequality leads to the following bound

[12]:

$$\mathrm{P}(\sup_{E} |\nu_l(E) - \mathrm{P}(E)| > t\,) \leqslant 4\, m_{\mathcal{E}}(2l)\, exp(-\tfrac{l}{8}\, t^2) \tag{2.3}$$

This inequality finally provides a sufficient condition for the uniform convergence of ν_l with respect to $\mathcal{E}$. Replacing the growth function with its upper bound as specified before, the right-hand side (eq. 2.3) does only converge for $l \to \infty$, if there exists $h_{\mathcal{E}}$ meaning that $m_{\mathcal{E}}(2l)$ is polynomially bounded in l for a sufficiently large l.

An even stronger result has been proven, which provides a sufficient as well necessary condition for the uniform convergence of empirical frequencies to their means. For this purpose, define $H_{\mathcal{E}}(l) := \mathbb{E}(ln(\kappa_{\mathcal{E}}(v^{(l)})))$, which is called the entropy of $\mathcal{E}$. Then, the uniform convergence of ν_l with respect to $\mathcal{E}$ is satisfied if and only if $\frac{H_{\mathcal{E}}(l)}{l} \to 0 \,\mathrm{as}\, l \to \infty$ [12]. The relation is usually of little practical relevance. But it provides an equivalent to the concept of uniform convergence and is therefore fundamental. It also verifies the previously derived sufficient condition. Because $H_{\mathcal{E}}(l) \leqslant ln(m_{\mathcal{E}}(l))$, the convergence of $\frac{ln(m_{\mathcal{E}}(l))}{l} \to 0$ implies uniform convergence. However, this is only satisfied if $m_{\mathcal{E}}(l)$ can not grow exponentially for arbitrary l.

Following these insights, the empirical risk estimator I_{emp} does uniformly converge to the risk functional, if the capacity of the set $\mathcal{E}_Q$ is finite. According to the declared properties of good estimators, the uniform convergence implies that the risk value of the empirical risk minimizer does arbitrarily closely approximate the risk functional's minimum (eq. 2.1). The capacity of $\mathcal{E}_Q$ is a measure of the complexity of the function space $\mathcal{F}$, because it quantifies the separation ability of a set of decision rules. To make this connection even more clear, one may consider $\mathcal{E}_{\mathcal{F}} := \{\{v \in \{0,1\} \times \mathbb{R}^d \,|\, f_\alpha(x) = 1\} \,|\, f_\alpha \in \mathcal{F}\}$. These events perfectly characterize the various decision rules. Obviously, the implied subsets on a given sample by $\mathcal{E}_Q$ and $\mathcal{E}_{\mathcal{F}}$ are identical. This is because two functions which classify the same objects differently than observed are also these, which map the same observations to class one, due to the binary classification. The same reasoning holds vice versa. Thus, the capacities of $\mathcal{E}_Q$ and $\mathcal{E}_{\mathcal{F}}$ are identical. Therefore, one defines $h_{\mathcal{F}} := h_{\mathcal{E}_Q}$ as the capacity of the set of decision rules $\mathcal{F}$. The derived bounds can furthermore be used to provide a confidence interval for the risk of the empirical risk minimizer $I(\alpha_{emp})$ in the canonical way mentioned before (sec. 2.2.2). To do so, one has to solve the right-hand side of inequality 2.3 for t after replacing $m_{\mathcal{E}}(2l)$ with its respective bound depending on l. It results $\varkappa(\eta, l, h_{\mathcal{F}})$, according to which $I_{emp}(\alpha) - \varkappa(\eta, l, h_{\mathcal{F}}) < I(\alpha) < I_{emp}(\alpha) + \varkappa(\eta, l, h_{\mathcal{F}})$ holds simultaneously for all α with a probability of at least η. Thus, for the empirical risk minimizer it follows: $I_{emp}(\alpha_{emp}) - \varkappa(\eta, l, h_{\mathcal{F}}) < I(\alpha_{emp}) < I_{emp}(\alpha) + \varkappa(\eta, l, h_{\mathcal{F}})$. Depending on the concrete assumptions made during the whole derivation, other more tight intervals are possible. This is why at this point a concrete interval is omitted. Instead, it is referred to [3, ch. 6.9, 6.10] for greater detail. However, all intervals have in common, that the precision depends increasingly on the quotient $\frac{h_{\mathcal{F}}}{l}$. Consequently, the higher the capacity the smaller is the uniform convergence's rate. Therefore, the bounds mirror the behaviour of the previously mentioned generalization-approximation trade-off. As on the one hand, the more powerful the chosen space $\mathcal{F}$ is, the smaller is the best possible risk value. On the other hand, due to an increased capacity the deviation $\varkappa(\eta, l, h_{\mathcal{F}})$ increases and the estimated risk minimizer does not necessarily represent a small risk value.

Regression estimation

Similar results about the uniform convergence can be obtained for the problem of interpreting direct experiments [9]. Because the risk functional is real-valued the derivation relies on a discretization of it. This allows to reduce the problem to the uniform convergence of frequencies to their probabilities of the previous section. Furthermore, because the loss function Q is not naturally bounded anymore, the existence of a finite $q := \sup\limits_{v,\alpha} Q(v, \alpha)$ has to be assumed. Central idea is to utilize the (Lebesgue's) integral definition to represent the risk and empirical risk functional by their limit series [3, ch. 7.2, 7.4]. Because $Q(., \alpha)$ is a positive random variable for all α, the following holds by definition of the integral operator:

$$I(\alpha) = \lim_{n\to\infty} \sum_{j=1}^{n} \frac{q}{n} \, \mathrm{P}(Q(v, \alpha) > \tfrac{j}{n} q) \quad I_{emp}(\alpha) = \lim_{n\to\infty} \sum_{j=1}^{n} \frac{q}{n} \, \nu_l(Q(v, \alpha) > \tfrac{j}{n} q) \tag{2.4}$$

as before $\nu_l(.)$ denotes the frequency of the respective event with respect to a sample $v^{(l)}$. Defining the events $E_{\alpha,j,n} := \{v \,|\, Q(v, \alpha) > \frac{j}{n} q,\ j, n \in \mathbb{N},\ j \leqslant n\}$ leads

[9] Or commonly known as regression estimation.

to the following bound for an arbitrary α:

$$|I(\alpha) - I_{emp}(\alpha)| \leqslant \lim_{n\to\infty} \sum_{j=1}^{n} \frac{q}{n} \left|\mathrm{P}(E_{\alpha,j,n}) - \nu_l(E_{\alpha,j,n})\right| \leqslant q \sup_{j,\,n} \left|\mathrm{P}(E_{\alpha,j,n}) - \nu_l(E_{\alpha,j,n})\right| \tag{2.5}$$

The supremum on the right-hand side already resembles the uniform convergence of frequencies of events to the respective probabilities. To generalize the set of events, one defines: $E_{\alpha,b} := \{v \,|\, Q(v,\alpha) > b,\, b \in \mathbb{R}^+\}$. Then, for an arbitrary α it holds: $\sup_{j,n} |\mathrm{P}(E_{\alpha,j,n}) - \nu_l(E_{\alpha,j,n})| \leqslant \sup_b |\mathrm{P}(E_{\alpha,b}) - \nu_l(E_{\alpha,b})|$. On the basis of eq. 2.5 it follows:

$$\begin{aligned} &|I(\alpha) - I_{emp}(\alpha)| \leqslant q \sup\nolimits_b |\mathrm{P}(E_{\alpha,b}) - \nu_l(E_{\alpha,b})| \\ \Longrightarrow\ &\mathrm{P}(\,\sup\nolimits_\alpha |I(\alpha) - I_{emp}(\alpha)| > t\,q\,) \leqslant \mathrm{P}(\,\sup\nolimits_{\alpha,b} |\mathrm{P}(E_{\alpha,b}) - \nu_l(E_{\alpha,b})| > t\,) \end{aligned} \tag{2.6}$$

Hence, if the right-hand side converges (eq. 2.6) the empirical risk functional does uniformly converge. Because the term on the right-hand side is itself a case of uniform convergence of frequencies of events, the results of the previous section (eq. 2.3) can be applied to bound this probability. It follows:

$$\begin{aligned} \mathrm{P}(\,\sup_\alpha |I(\alpha) - I_{emp}(\alpha)|\, > t\,q\,) &\leqslant \mathrm{P}(\sup_{\alpha,b} |\mathrm{P}(E_{\alpha,b}) - \nu_l(E_{\alpha,b})| > t\,) \\ &\leqslant 4\, m_{\mathcal{E}}(2\,l)\, exp(-\tfrac{l}{8}\, t^2) \end{aligned} \tag{2.7}$$

with $\mathcal{E} = \{E_{\alpha,b} \,|\, f_\alpha \in \mathcal{F},\, b \in \mathbb{R}^+\}$ and l being sufficiently large [10]. Analogue to the pattern recognition problem, if the set $\mathcal{E}$ has finite capacity, the growth function is polynomially bounded and the convergence is satisfied. The capacity of $\mathcal{E}_\mathcal{F} = \{\{v \,|\, Q(v,\alpha) - b > 0\} \,|\, f_\alpha \in \mathcal{F},\, b \in \mathbb{R}^+\}$ is the defined capacity $h_\mathcal{F}$ of the function space $\mathcal{F}$ in the context of the regression estimation problem [3, ch. 7.4]. It is easily seen, that this encompasses the given definition for the pattern recognition problem, because in this case Q can only take on the values 0 or 1. $\mathcal{E}_\mathcal{F}$ characterizes the functions in $\mathcal{F}$ in terms of regions of equivalent prediction quality. The subsets implied by $\mathcal{E}_\mathcal{F}$ on a given sample $v^{(l)}$ in terms of table 2.1 can be imagined in the following way. A function f_α does assign to each observation v_i a different loss value $Q(v_i,\alpha)$. It can be assumed that the loss values are different for each v_i as otherwise a sample is chosen, such that this is satisfied. Then, an order is implied

[10] According to the derivation in sec. 2.2.3 the bound holds for $l > \frac{2}{t^2}$. Furthermore, the growth function is only polynomially bounded, if $l > h_\mathcal{E}$.

to the v_i, which is $v_{j_1} > v_{j_2} \iff Q(v_{j_1}, \alpha) > Q(v_{j_2}, \alpha)$. This "lines up" the l observations: $v_{j_1} > ... > v_{j_l}$ and $Q(v_{j_1}, \alpha) > ... > Q(v_{j_l}, \alpha)$. The values $Q(v_{j_k}, \alpha)$ are exactly the levels $b_{\alpha,k}$ which define the sets associated to f_α in $\mathcal{E}_\mathcal{F}$, which contain different subsets of the given sample. Each function refers to l subsets of the sample, which can be separated by $\mathcal{E}_\mathcal{F}$. These are $\{v_{j_1}\}, \{v_{j_1}, v_{j_2}\}, \{v_{j_1}, v_{j_2}, v_{j_3}\}, ...$. In total $l! > 2^l$ permutations are possible. The capacity of $\mathcal{F}$ measures, if there are enough functions available to produce sufficiently many permutations to cover all possible subsets of observations.

Regarding the bounds of the risk functional, it might be meaningful, to allow bigger deviations depending on the magnitude of the risk value. Therefore, the SL-theory provides an extended analysis of the quantity $\sup_\alpha \frac{|I(\alpha) - I_{emp}(\alpha)|}{I(\alpha)}$ [3, ch. 7.5]. Following this notion leads to improved confidence intervals. The uniform convergence of the relative empirical risk does still imply the uniform convergence of the empirical risk. This is because $\frac{\sup_\alpha |I(\alpha) - I_{emp}(\alpha)|}{\sup_\alpha I(\alpha)} \leqslant \frac{\sup_\alpha |I(\alpha) - I_{emp}(\alpha)|}{I(\alpha)}$. The analysis of the relative risk bases on assumptions about the bound of the relative p^{th} order mean of the loss, which is $\frac{\sqrt[p]{\mathbb{E}(Q(V,\alpha)^{2p}))}}{\mathbb{E}(Q(V,\alpha)^2)}$ [11]. The analysis shows that the highest convergence rate results, if the boundedness for $p = 2$ is given. Furthermore, a sufficient and necessary condition for uniform convergence in the regression estimation problem exists as well. Explicit derivations can be found in [3, ch. 7.A].

2.2.4 Structural risk minimization

The conducted analysis of the two problems of statistical dependency estimation, has produced an uniform bound of I in terms of the empirical risk I_{emp}, the sample size l, the growth function $m_{\mathcal{E}_\mathcal{F}}$ and the cover probability η. This relation allows to derive a sufficient condition under which the method of empirical risk minimization is justified, in terms of convergence and calculable confidence intervals. Assuming the existence of a finite capacity $h_\mathcal{F}$ and a sufficiently large sample size l, the uniform deviation between I and I_{emp} is denoted by $\varkappa(\eta, l, h_\mathcal{F})$. Although different concrete bounds exist, depending on whether one analyses the relative risk or what assumptions about the distribution of $Q(., \alpha)$ are made, all have in

[11] For $p = 2$ the requirement is equivalent to the uniform boundedness of the relative variance.

common that $\varkappa$ depends monotonically increasing on the quotient $\frac{h_{\mathcal{F}}}{l}$ [3, ch. 8]. Thus, if there is access to an arbitrary number l of observations, the empirical risk minimizer α_{emp} could be chosen, such that the associated risk arbitrarily closely approximates the real minimum (see eq. 2.1). Under this assumption, the function space, which admits the smaller minimum of risk $\min\limits_{f_\alpha \in \mathcal{F}} I(\alpha)$ is to be preferred, as long as $h_{\mathcal{F}}$ is finite. Hence, increasing complexity will only improve the estimation. However, in practice any sample size is limited. The risk of the empirical risk minimizer among the more complex function space is not necessarily smaller than for a more restricted space due to the increased offset $\varkappa$, because of the higher capacity. This is established by the inequality $I(\alpha_{emp}) < I_{emp}(\alpha_{emp}) + \varkappa(\eta, l, h_{\mathcal{F}})$. The trade-off shows, that the pure empirical risk minimization is incomplete. Instead, it is required to control the capacity. Based on these thoughts the SL-theory proposes an alternative, which is the structural risk minimization (SRM). Instead of solely minimizing the empirical risk, the upper bound of the risk functional $I(\alpha) < I_{emp}(\alpha) + \varkappa(\eta, l, h_{\mathcal{F}})$ is minimized. This requires a set of function spaces over which is to be optimized. For this purpose, consider a space of functions $\mathcal{F}$ with a finite capacity. This space is decomposed into a finite sequence of nested subsets called a structure [12]: $S_1 \subset S_2 \subset ... \subset S_q$. The structure has to be defined a priori independently of the observations. The structure defines multiple (q) function spaces, of increasing capacity, due to the inclusions: $h_1 \leqslant h_2 \leqslant ... \leqslant h_q$. The SRM estimator is the minimizer of the empirical upper bound estimate: $h_{SRM}, \alpha_{SRM} := \operatorname*{argmin}\limits_{h_i, \alpha: f_\alpha \in S_i} I_{emp}(\alpha) + \varkappa(\eta, l, h_i)$. Within any given subset S_i, the upper bound is minimized by the respective empirical risk minimizer. For the risk of the structural risk minimizer the bound $I(a_{SRM}) < I_{emp}(\alpha_{SRM}) + \varkappa(\eta, l, h_{SRM})$ is valid with a probability of at least $1 - q\,\eta$. Therefore, the confidence probability η could be adjusted to $\frac{\eta}{q}$. Although the increase in the bound $\varkappa$ is for $q < 100$ only slightly bigger [3, ch. 8], it shows, that the more extensive the structure is, the worse does the SRM generalize. The SRM method essentially consists of two stages. The first lies in properly defining the structure. The subset relation reflects in some sense the prior knowledge about the desired dependency. Functions which are more likely (a priori) to approximate well the desired dependency should be

[12] A common example of a structure are polynomials grouped by their degree.

gathered in lower indexed subsets. This is because these obtain some kind of credit, due to the smaller capacity. The second component is the proper estimator of the risk value of the empirical risk minimizer. One option are clearly the derived upper bounds, obtained by the previous analysis of the uniform convergence. However, in principle different estimators are admissible too. An alternative is the so called moving-control estimator, which is commonly known as leave-one-out cross validation [3, ch. 8.1]. It estimates the expected risk $\mathbb{E}(I(\alpha_{emp}))$ and can alternatively be used to select a function within the structure, by testing successively $\alpha_{emp,i}$, i.e. the empirical risk minimizer within S_i. Note, that this expectation of the risk respects the randomness in the estimation, due to the random sample $v^{(l)}$. If this expectation was known, then the same on-expectation-best estimator would be chosen, independent of the given sample. However, the value of $I_{emp}(\alpha_{emp}) + \varkappa(\eta, l, h)$ is supposed to bound explicitly the random variable $I(\alpha_{emp})$. Therefore, applying this bound intends to find the sample wise optimum. Hence, even if the risk of the estimated function was known, the selection among the estimators of the structure would be sample dependent. In [3, ch. 8] multiple examples of possible structures based on different function spaces and problems are given. The SRM method is also the foundation of the well-known SVM classifier. Furthermore, based on the results of the uniform convergence and the SRM method an analysis of the third kind of introduced dependencies, which are inverse problems is conducted. Details can be found in [3, ch. 10].

2.3 Reproducing kernel Hilbert space

So far different kinds of problems such as ill posed problems or the statistical estimation of dependencies have been presented together with an analysis of how to solve them. In this section a specific function space shall be introduced, which is often utilized in statistical estimation problems. This is the reproducing kernel Hilbert space (RKHS). Besides the general definition, different properties are presented, which are essential in the context of the SL-theory. The information bases on the work of Cucker [14] and Paulsen [15]. In the appendix a collection of basic definitions of functional analysis is provided. A comprehensive introduction into functional analysis can be obtained in [16].

2.3.1 Definition

Let $\mathcal{H}$ be a Hilbert space of complex valued functions over a domain $\mathcal{X}$, equipped with the common pointwise addition and scalar multiplication. Mostly $\mathcal{X}$ is a compact subset of the $\mathbb{R}^n$. The evaluation functional L_x for $x \in \mathcal{X}$, is a linear functional, which maps each function f of $\mathcal{H}$ to its evaluation at the point x. This means, $L_x : \mathcal{H} \to \mathbb{C}, f \to L_x(f) = f(x)$. $\mathcal{H}$ is a RKHS, if for all x in $\mathcal{X}$ the respective evaluation functional is continuous. Due to the linearity of L_x, this is equivalent to being a bounded operator, i.e. $\|L_x(f)\| = |f(x)| \leqslant R_x \|f\|_{\mathcal{H}}$ for all $f \in \mathcal{H}$ and some $R_x \in \mathbb{C}$. From the boundedness of L_x follows, that convergence in $\mathcal{H}$ implies pointwise convergence. This immediately explains, why for example the L^2 space is not a RKHS. Note, that in the definition the uniform boundedness of L_x over $\mathcal{X}$ is not required. However, if this is given, that is R_x is bounded over x, then uniform convergence is implied by convergence in $\mathcal{H}$.

2.3.2 Kernel of a RKHS

The central concept of the RKHS is the associated kernel. This gives a convenient representation and allows to characterize any RKHS. The Riesz-Frechet representation theorem (app.A) states that for each bounded, linear functional L of the dual space $\mathcal{H}'$ of any Hilbert space $\mathcal{H}$, there exists an unique element g_L of $\mathcal{H}$, such that the two properties are satisfied: 1) $L(f) = \langle f, g_L \rangle$ for all f in $\mathcal{H}$ 2) $\|L\| = \|g_L\|_{\mathcal{H}}$. The associated function $\mathcal{H} \to \mathcal{H}'$ is a conjugate linear, isometric isomorphism [16, p.246]. Therefore, in case of the RKHS it is possible to apply the theorem to the evaluation functionals, as these are bounded by definition. Thus, for each L_x there is a function K_x of $\mathcal{H}$, such that $L_x(f) = f(x) = \langle f, K_x \rangle$. Because K_x itself is element of $\mathcal{H}$, it holds $K_x(y) = L_y(K_x) = \langle K_x, K_y \rangle$ for any y in $\mathcal{X}$. This relation leads to following definition:

$$K_{\mathcal{H}} : \mathcal{X} \times \mathcal{X} \to \mathbb{C} : K_{\mathcal{H}}(x, y) = \langle K_x, K_y \rangle_{\mathcal{H}} \tag{2.8}$$

The function $K_{\mathcal{H}}$ is called the reproducing kernel of $\mathcal{H}$. It holds $K_x(y) = K_{\mathcal{H}}(x, y)$. Some of the most relevant properties of a kernel shall be mentioned:

- $K_{\mathcal{H}}$ is conjugate symmetric: $K_{\mathcal{H}}(x, y) = \langle K_x, K_y \rangle = \overline{\langle K_y, K_x \rangle} = \overline{K_{\mathcal{H}}(y, x)}$.

- $K_{\mathcal{H}}$ is a positive semi-definite function. A function f is called positive (semi-) definite, if for any finite sequence $(x_1, ..., x_n)$ the matrix $m_{i,j} = f(x_i, x_j)$ is positive (semi-) definite [15, 14]. For the kernel $K_{\mathcal{H}}$ the definiteness results, because all possible matrices are Gram matrices.

- The kernel function $K_{\mathcal{H}}$ is not necessarily continuous. However, if $K_{\mathcal{H}}$ is continuous then the space $\mathcal{H}$ consists of continuous functions [14]. This follows by:
  - $\mathrm{span}(\{K_x \,|\, x \in \mathcal{X}\})$ is dense in $\mathcal{H}$ [13]
  - $|f(x)| = |L_x(f)| = |\langle f, K_x\rangle| \leqslant \sqrt{|\langle f, f\rangle|}\sqrt{\langle K_x, K_x\rangle} = \|f\|_{\mathcal{H}} \|K_x\|_{\mathcal{H}} = \|f\|_{\mathcal{H}} \sqrt{K(x,x)}$. Assuming the existence of a finite supremum $C_K := \sup_{x,y} |K_{\mathcal{H}}(x,y)|$, it follows: $\|f\|_\infty \leqslant \|f\|_{\mathcal{H}} \sqrt{C_K}$. Note, if $K_{\mathcal{H}}$ is continuous and $\mathcal{X}$ compact C_K exists. This shows, that the convergence in $\mathcal{H}$ implies the convergence in $\mathrm{C}(\mathcal{X})$. If K is continuous, then the functions $K_x = K_{\mathcal{H}}(.,x)$ are continuous too. Due to the dense subspace each function of $\mathcal{H}$ is therefore a limit of continuous functions. Because continuity is preserved by uniform convergence the proposition follows.

If a kernel is additionally continuous, it is called a Mercer kernel.

2.3.3 Moore-Aronszajn's theorem

It is intuitive, that the kernel of a RKHS does characterize the space quite specifically, as it is intrinsically connected to its elements. This raises the question, how tight the relation between a RKHS and the implied kernel is. Especially whether two different RKHSs can have the same kernel. This is matter of the theorem by Moore-Aronszajn, which leads to the conclusion, that each RKHS is uniquely identified by its kernel. In general a function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{C}$ is called a kernel function, if it is conjugate symmetric and positive semi-definite. Let $K_{\mathcal{H}}$ denote the implied kernel of a RKHS $\mathcal{H}$. If K is a kernel function, then there exists a unique RKHS $\mathcal{H}_K$, such that $K_{\mathcal{H}_K} = K$. The proof is only outlined at this point. At first it is to show that there exists a RKHS with K as kernel. This space is constructed as follows: Consider the linear span of the functions $K_x := K(.,x)$ for x in $\mathcal{X}$. For this

[13] Due to the implication: $(\,(\forall g \in \mathcal{S}\,\langle f, g\rangle = 0) \implies f = 0\,) \implies \mathcal{S}$ dense in $\mathcal{H}$.

space, an inner product is defined by $\langle f, g\rangle := \sum_{i=1}^{n} \sum_{j=1}^{m} a_i \, \overline{b_j} \, K(x_i, y_i)$, at which $f = \sum_{i=1}^{n} a_i \, K_{x_i}$, $g = \sum_{j=1}^{m} b_j \, K_{x_j}$. Both together define an inner product space but not yet a Hilbert space. The desired Hilbert space results as the completion of this space [14]. It can be verified, that the resulting space is in fact a RKHS with K as kernel. The second part consists in proving, that this is the only RKHS with K as kernel. For this purpose one can show: If $\mathcal{H}_1, \mathcal{H}_2$ are two RKHS with kernels $K_{\mathcal{H}_1}, K_{\mathcal{H}_2}$, then $K_{\mathcal{H}_1} = K_{\mathcal{H}_2}$ implies equivalence $\mathcal{H}_1 = \mathcal{H}_2$ as well $\|f\|_1 = \|f\|_2$ for all functions. The proof is shown in [15].

This theorem is important, as it allows to completely define any RKHS by defining a proper kernel function, which is for practical applications quite convenient. Proving the symmetry or continuity for a given function is rather simple, but proving it being positive (semi-) definite is rather complicated. However, the kernel property is invariant under certain operations such as addition and others, which is why kernels can be constructed recursively.

2.3.4 Mercer's theorem

Mercer's theorem is a sentence in the functional analysis and operates outside the theory of RKHSs. The utilization of the theorem regarding RKHSs allows to further characterize the elements of the space by means of the kernel function. This representation enables a different perspective and is also the basis for further proofs regarding properties of RKHSs. The theorem involves some advanced concepts of functional analysis, which is why a complete proof is out of the scope of this thesis. Furthermore, the literature is rather inconsistent in presenting its application to the theory of RKHS, which is why the theorem and its use in RKHS is only outlined. The given formulation is from [16, p.301], which also provides a full proof of the theorem:

Let K be a complex-valued function of $\mathrm{C}([0,1]^2)$ and $T_K : L^2([0,1]) \to L^2([0,1])$ be the corresponding integral operator, $(T_K(f))(t) = \int_0^1 K(s,t) \, f(s) \, \mathrm{d}s$. Let $K(s,t) = \overline{K(t,s)}$ for $s, t \in [0,1]$ which implies T_K being self-adjoint (app.A). Let $\lambda_1, \lambda_2, ...$ be the eigenvalues different than 0 of T_K counted by their geometric number and $e_1, e_2, ...$ the corresponding eigenfunctions. The eigenfunctions build an orthonor-

[14] For the definition of the completion of a normed vector space see app.A.

mal basis in $L^2([0,1])$. If T_K is positive definite, then $K(s,t) = \sum_{j=1}^{\infty} \lambda_j \, e_j(s) \, \overline{e_j(t)}$, at which the convergence is absolute and uniform.

Note, that the integral operator maps into the subspace of continuous functions, however the inclusion into the Hilbert space L^2 is considered. An operator T is called self-adjoint if it is its adjoint operator $T^* = T$. A linear operator $T : \mathcal{H} \to \mathcal{H}$ is called positive, if it is self-adjoint and $\forall x \in \mathcal{H} : \langle Tx, x \rangle \geqslant 0$. Absolute convergence is equivalent to the known definition of finite dimensional spaces. However, different to the finite dimensional case, it does not imply the equivalence of any ordering in the series. The theorem applies the spectral theory of compact linear operators of Hilbert spaces here $L^2([0,1])$. An even more special case is faced, due to T_K being self-adjoint and positive. This theory is convenient in the sense, that it admits many results similar to the finite dimensional case. For example the number of eigenvalues different than zero is at most countable. Furthermore, all eigenvalues are non-negative, real values due to the positivity of the operator [16, p.294]. Also, the eigenfunctions of positive eigenvalues are continuous [14]. The continuity of the operator T_K follows by the continuity of the function K. The required compactness can be shown by means of Arzela-Ascoli's theorem [16, p.73]. The main aspect of the theorem is the uniform and absolute convergence, which specifically comes from the positivity. According to [14] the theorem is also valid in a more general context. For example, in case of a compact set $\mathcal{X}$ instead of $[0,1]^2$ or when the integral is defined with respect to a measure μ.

Consider a Mercer kernel K. Then all requirements of Mercer's theorem are satisfied and it is applicable to the function K. This allows to characterize the functions of the associated RKHS $\mathcal{H}_K$ and its inner product through the eigenfunctions and the inner product of the $L^2_\mu(\mathcal{X})$ space. For this purpose, define the space $H = \{f \in L^2_{\mu(\mathcal{X})} \,|\, f = \sum_{k=1}^{\infty} a_k \, e_k$, with $(\frac{a_k}{\sqrt{\lambda_k}}) \in l^2$, $a_k \in \mathbb{C}\}$, with convergence in $L^2_{\mu(\mathcal{X})}$ and λ_k, e_k being the respective eigenvalues and eigenfunctions according to Mercer's theorem. On this space an inner product is defined by $\langle f, g \rangle_{\mathcal{H}} := \sum_{k=1}^{\infty} \frac{a_k \, \bar{b}_k}{\lambda_k}$. Both together define a RKHS and it can be verified, that this is exactly the RKHS $\mathcal{H}_K$ associated to the kernel function K [14].

2.3.5 Feature maps

The concept of feature maps is used to put an additional intuition behind the utilization of kernels. A feature map is a function $\phi : \mathcal{X} \to \mathcal{G}$, where $\mathcal{G}$ is a Hilbert space. Usually $\mathcal{X}$ is the sample space of an estimation problem and describes the set of observable features. The feature map is a transformation of each (finite-dimensional) feature to a more evolved concept. There are different ways to define a RKHS on the basis of a feature map. A canonical way is to define the kernel $K_\phi(x, y) := \langle \phi(x), \phi(y) \rangle_{\mathcal{G}}$, which implies the RKHS $\mathcal{H}_{K_\phi}$ as described before. On the other hand, for each RKHS $\mathcal{H}_K$ there are (multiple) feature maps ϕ_K, such that $H_{K_{\phi_K}} = H_K$. A canonically example was already presented. This is given by the function, $\phi : x \to K_x = K(., x)$. Mercer's theorem gives rise to a different feature map, which is $\phi : \mathcal{X} \to l^2$, $\phi(x) = (\sqrt{\lambda_i}\, e_i(x))_i$. The function ϕ is indeed well-defined, continuous and satisfies $K(x, y) = \langle \phi(x), \phi(y) \rangle_{l^2}$ [14]. The notion of the feature map suggests the interpretation of the functions K_x as rather simple concepts, e.g. the scalar product of l^2, on top of a complex feature transformation. In other words, the actual function is simple, if the correct perspective regarding the features is perceived. Thus, the complexity of the RKHS reduces to the complexity of a proper feature transformation.

2.3.6 Representer theorem

A property, which makes the RKHS a useful function space for statistical estimation problems as introduced in the SL-theory is the representer theorem. So far it was shown, that in order to define a RKHS it suffices to specify a kernel function. Although the implied functions can be characterized by means of Mercer's theorem, this remains a rather theoretical insight and is not very practical for tasks like optimization. However, the representer theorem reveals that the empirical risk minimizer among a RKHS has a well-formed representation, which makes the actual calculation very convenient. Consider the following optimization problem:

$$\min_{f \in \mathcal{H}_K} \sum_{i=1}^{l} Q(v_i, f) + o(\|f\|_{\mathcal{H}_K}) \tag{2.9}$$

at which o is a monotonically increasing real-valued function, K a kernel function, Q a loss function and $v_i = (x_i, y_i)$, as introduced before in the context of the SL-theory. Then the minimizing function f^* within a RKHS $\mathcal{H}_K$ is of the form: $f^* = \sum_{i=1}^{l} a_i K(., x_i) \; a_i \in \mathbb{C}$. Therefore, the solution can be represented as a liner combination in terms of the kernel's partial functions at the points x_i. Hence, by replacing this representation in the original optimization problem the minimizer can actually be calculated. The optimization problem (eq.2.9) comprises an even more general problem set than the initial empirical risk minimization presented before. The proof of the theorem uses the characterization according to Mercer's theorem and is shown in [14].

3 Rethinking statistical learning theory: LUSI

In 2018 Vapnik and Izmailov have released a paper about "Rethinking statistical learning theory: Learning using statistical invariants (LUSI)" [11]. In their work a new method for the pattern recognition problem was introduced. The basic procedure aims for estimating the conditional probability function of the data generating distribution, as means to enable classification in a canonical way through probability information about the respective class. The method directly applies the aforementioned Tikhonov regularization principle and considers a RKHS for approximation. As a result, it arises a quadratically optimization problem. To enhance the estimation with further information so called statistical invariants are defined and added as linear side constraints. The estimator is the minimization of the resulting optimization problem. In the following chapter the method will precisely be defined and discussed.

3.1 Definition

Conditional probability function

The context is described by the setting of the pattern recognition problem (sec. 2.2.1). Accordingly, the available information is in the form of a random i.i.d sample $(y_i, x_i)^{(l)}$. It is assumed, that x is from a (compact) domain $\mathcal{X} = [0, C]$ of $\mathbb{R}^d$. Although in practice this might be achievable only through certain transformations, this is not really a tight restriction but simplifies calculations. The object of interest is the function $\lambda(x) := \mathrm{P}(Y = 1 \,|\, x)$, i.e. the conditional probability of observing Y being 1 having observed the representative x. Based on this information a decision rule can be derived in a canonical way, which is $r(x) = \mathbb{1}_{\{\lambda(x) > 0.5\}}(x)$.

The desired conditional probability function can formally be defined as the function satisfying the following integral equation among the Lebesgue integrable functions with respect to the measure $\mathrm{P}(x)$, mapping into $[0, 1]$:

$$\int_0^{\tilde{x}} f(x)\,\mathrm{dP}(x) = \mathrm{P}(Y = 1, X < \tilde{x}) =: F_1(\tilde{x}) \tag{3.1}$$

Plugging in λ for f reveals its admissibility. The integral definition is advantageous as it does not require the existence of any density functions to define the conditional probability as their respective quotient. The definition of the conditional probability function in terms of functions of type F_1 implies an inverse problem in the form of a linear operator equation. Under concrete specifications of the integral operator, i.e. $\mathrm{P}(x)$, as well as the function spaces encompassing λ and F_1 respectively, the associated direct relation can be explicitly inferred. For example, one might assume continuity of λ together with a uniform distribution over $\mathcal{X}$. Then the conditional probability function is obtained by applying the inverse integral operator, which means differentiation to the distribution alike function on the right-hand side. Although this is a rather special case, it reveals that even under these assumptions the problem is ill-posed. This is because the differentiating operator is unbounded and therefore not continuous as has been proven before (sec. 2.1.1). Note, that the functions used in this proof were also distribution functions. Therefore, even though the space of functions can be restricted from the general set of differentiable functions to the set of differentiable distribution functions, the differentiating operator remains discontinuous [1]. Consequently, the calculation of the conditional probability function based on the stated inverse problem is ill-posed. Because the function F_1 is unknown and can only be estimated the ill-posed property causes a special difficulty as explained. However, a framework how to "solve" these kinds of problems has been already introduced. This is the regularization principle by Tikhonov (sec. 2.1.2). According to the theory the solution space $\mathcal{Z}$ becomes the set of possible conditional probability functions, while the data space $\mathcal{G}$ is the set of approximations of the functions F_1 according to the chosen estimator of it. The associated operator A is the integral operator. To estimate F_1 on the basis of a given sample, the following relation is considered $\mathrm{P}(Y = 1, X < x) = \mathbb{E}(\mathbb{1}_{[0,x]}(X)\,Y)$,

[1] This should be related to the statement about restricting the data and solution space of ill-posed problems in sec. 2.1.1.

leading to the corresponding average estimate $\hat{F}_1(x) = \frac{1}{l}\sum_{i=1}^{l} y_i \, \mathbb{1}_{[0,x]}(x_i)$. The theorem by Clivenko-Cantelli gives additional justification for using this kind of non-parametric estimator. The given problem falls into an even more general case. Due to the unknown distribution $\mathrm{P}(x)$ both sides of the operator equation can only be approximated. For the estimation of the operator, the integral is rewritten as follows: $\int_0^{\tilde{x}} f(x)\,\mathrm{dP}(x) = \int_{\mathcal{X}} \mathbb{1}_{[0,\tilde{x}]}(x)\, f(x)\,\mathrm{dP}(x) = \mathbb{E}(\mathbb{1}_{[0,\tilde{x}]}(X)\, f(X))$. Hence, for a fixed function f the laws of large numbers give rise to the canonical average estimator $(\hat{A}f)(x) = \frac{1}{l}\sum_{i=1}^{l} \mathbb{1}_{[0,x]}(x_i)\, f(x_i)$. The set of admissible solutions $\mathcal{Z}$ is restricted to a function space $\mathcal{F}$. The authors decided to choose $\mathcal{F} = \{f = g + c \mid g \in \mathcal{H}_K,\ c \in \mathbb{R}\}$, where $\mathcal{H}_K$ is a RKHS of a selected kernel function K. $\mathcal{F}$ is therefore the union of all affine vector spaces regarding $\mathcal{H}_K$. The offset c does not add much expression, however the amount of constant functions is not necessarily subset of $\mathcal{H}_K$ [2]. For the data space $\mathcal{G}$ a reasonable choice is $L^2(\mathcal{X})$ (or $L^2_\mu(\mathcal{X})$), since both the function F_1 as well as its estimates fall inside and a proper Hilbert space is defined. Finally, Tikhonov's regularization operator can be constructed, leading to the associated approximate solution. According to the previous chapter, it follows (sec. 2.1.2):

$$\hat{\lambda} = \operatorname*{argmin}_{f \in \mathcal{F}} T_\gamma(\hat{A}f, \hat{F}_1)$$

$$T_\gamma(\hat{A}f, \hat{F}_1) := \int_{\mathcal{X}} ((\hat{A}f)(x) - \hat{F}_1(x))^2 \,\mathrm{d}x \;+\; \gamma\,\|f\|_{\mathcal{F}} \tag{3.2}$$

Substituting the corresponding estimators, the functional is reduced to the following form. Subsequently the term $\frac{1}{l}$ of any averages is ignored for obvious reasons. Considering the first part:

$$\begin{aligned}\int_{\mathcal{X}} ((\hat{A}f)(x) - \hat{F}_1(x))^2 \,\mathrm{d}x &= \sum_{i,j=1}^{l} f(x_i)\, f(x_j)\, V(i,j) - 2\sum_{i,j=1}^{l} f(x_i)\, y_j\, V(i,j) \;+\; \sum_{i,j=1}^{l} y_i\, y_j\, V(i,j) \\ &= \sum_{i,j=1}^{l} (y_i - f(x_i))\,(y_j - f(x_j))\, V(i,j) \end{aligned} \tag{3.3}$$

Here $V(i,j) := \int_{\mathcal{X}} \mathbb{1}_{[0,x_i]}(x)\, \mathbb{1}_{[0,x_j]}(x)\,\mathrm{d}x$. Because $\mathcal{X} = [0, C]$, $C = (c_1, ..., c_d) \in \mathbb{R}^d$, the expression results to $V(i,j) = \prod_{k=1}^{d} c_k - \max(x_{i_k}, x_{j_k})$, where x_{i_k} denotes the k^{th} coordinate. Concerning the norm term in equation 3.2 a slight adaption is made.

[2] The decision for an offset is rather arbitrary and should not confuse the reader.

Here $\|f = g + c\|$ is replaced by $\|g\|_{\mathcal{H}_K}$. Note, that this is not a real norm over $\mathcal{F}$ though. However, all originally made statements about the convergence of the Tikhonov approach remain valid. To actually calculate the norm the following is recognized. Although the functional T is not identical with the regularized empirical risk as formulated in the previous chapter (sec. 2.3.6), the representer theorem is still valid. Accordingly, the minimizing function is from $\mathrm{Span}(\{K_{x_1}, ..., K_{x_l}\}) + c$ where $K_{x_i}(x) = K(x, x_i)$. In the previous chapter, it was shown, how the RKHS is constructed on the basis of the kernel function K (sec. 2.3.3). For functions being a linear combination of K_x the scalar product and hence the norm is representable by a finite expression. It holds $\|g\|_{\mathcal{H}_K} = \sum\limits_{i,j=1}^{l} \alpha_i \, \alpha_j \, K(x_i, x_j)$, at which α_i are the respective coefficients of the linear combination of g. The desired conditional probability function has a bounded image between 0 and 1. Therefore, not all functions of the RKHS are meaningful. Only functions, which satisfy this condition are of actual interest. For this purpose, one would need to add the constraint $0 \leqslant f(x) \leqslant 1 \, \forall \, x \in \mathcal{X}$ to the optimization problem. Because this is practically not controllable, this property is only demanded at the observations x_i.

Statistical invariants

The so far introduced approach constitutes the first part of the LUSI method. It remains to incorporate the idea of the statistical invariants. In the paper a "predicate" ψ is any function of the space $L^2_{\mathrm{P}(x)}$. The idea of statistical invariants consists in adding constraints to the optimization problem in the form $\langle f, \psi \rangle = \langle \lambda, \psi \rangle$, where $\langle ., . \rangle$ denotes the respective scalar product [3]. This kind of equation is called an "invariant". According to the aforementioned Riesz-Frechet representation theorem, each such predicate does define a linear functional over $L^2_{\mathrm{P}(x)}$. Thus, an invariant represents the demand of equivalence of the approximation f and the desired function λ regarding a linear functional of the dual space. However, there are two challenges. First, it is necessary to make a selection of a finite number of predicates, in order to actually control the invariants in the optimization problem. Secondly, the values of both parts of any invariant are unknown.

[3] Because $\mathcal{X}$ is a compact interval and λ is bounded, it is element of $L^2_{\mathrm{P}(x)}$. Hence the linear functional is defined. The same usually holds for $f \in \mathcal{H}_K$.

As alternative the invariants are replaced by equations, which result by substituting the left- and right-hand side with appropriate estimates. For $f,\, \psi \in L^2_{\mathrm{P}(X)}$ it holds $\langle f, \psi \rangle_{L^2_{\mathrm{P}(x)}} = \int_{\mathcal{X}} f(x)\, \psi(x)\, \mathrm{dP}(x)$, thus it can canonically be estimated by the empirical average $\frac{1}{l} \sum_{i=1}^{l} f(x_i)\, \psi(x_i)$. Since λ is unknown, the following relation is utilized $\langle \lambda, \psi \rangle_{L^2_{\mathrm{P}(x)}} = \int_{\mathcal{X}} \mathrm{P}(Y = 1 \,|\, x)\, \psi(x)\, \mathrm{dP}(x) = \mathbb{E}(Y\, \psi(X))$, enabling the estimation by $\frac{1}{l} \sum_{i=1}^{l} y_i\, \psi(x_i)$. Therefore, for a set of m predicates $\psi_1, ..., \psi_m$ the corresponding invariants are replaced by the statistical alternatives. The resulting m equations are added as additional, linear side constraints to the optimization problem (eq. 3.2). Thus, only functions are considered, whose estimated predicate property equals the corresponding observed value of the conditional probability function. Bringing all together the LUSI estimate results as minimizer of the following optimization problem. Introducing a matrix notation this is:

$$\begin{aligned} \min_{\alpha, c} \quad & (\mathbb{K}\, \alpha + c\, 1_l)^\intercal\, \mathbb{V}\, (\mathbb{K}\, \alpha + c\, 1_l) - 2\, (\mathbb{K}\, \alpha + c\, 1_l)^\intercal\, \mathbb{V}\, y + \gamma\, \alpha^\intercal \mathbb{K}\, \alpha \\ \text{s.t.} \quad & \Psi\, (\mathbb{K}\, \alpha + c\, 1_l) = \Psi\, y \\ & 0_l \leqslant \mathbb{K}\, \alpha + c\, 1_l \leqslant 1_l \end{aligned} \tag{3.4}$$

here $\mathbb{K},\, \mathbb{V} \in \mathbb{R}^{l \times l}$ with $\mathbb{K}_{i,j} = K(x_i, x_j)$, $\mathbb{V}_{i,j} = V(i,j) = V(j,i) = \mathbb{V}_{j,i}$, furthermore $\Psi \in \mathbb{R}^{m \times l}$, with $\Psi_{i,j} = \psi_i(x_j)$ for $0 \leqslant i \leqslant m, 0 \leqslant j \leqslant l$, as well as $\alpha,\, y, 0_l, 1_l \in \mathbb{R}^l$, with $\alpha = (\alpha_1, ..., \alpha_l)$, $y^{(l)} = (y_1, ..., y_l)$, $1^{(l)} = (1, ..., 1)$. The term $\frac{1}{l}$ cancels out for the side constraints.

3.2 Discussion

3.2.1 Why Rethinking?

In the published paper [11], the LUSI method was introduced as a "Rethinking" of statistical learning theory. In the following this aspect shall be investigated in a little bit more detail, explaining where the change in paradigm actually lies. As argued in the theory chapter, three basic tasks are differentiated in the context of statistical dependency estimation. The overarching purpose of this method does coincide with the pattern recognition task. For this, the objective was declared to attain a certain quality when classifying objects according to a decision rule r, usually expressed by the risk functional $\int_{\mathcal{Y} \times \mathcal{X}} (y - r(x))^2\, \mathrm{dP}(y, x)$ (sec. 2.2.1). The risk functional should be optimized among a specified set of decision rules. Note, that this only defines the objective, but not necessarily the strategy how to achieve it. One option to do so, is the empirical risk minimization. However, the intrinsic focus is not identifying any "true" function, but to aim for a minimal probability of incorrect classification. This is also why an arbitrary type of decision rules could be chosen for minimization. Hence, why should the explicit estimation of the conditional probability function still be a reasonable idea to minimize this risk functional? First, considering the definition of the Bayes error, the best decision rule in the sense of the this risk functional is defined by: $r^*(x) := \mathbb{1}_{\{\lambda(x)>0.5\}}(x)$. Thus, the conditional probability function λ gives the complete information to obtain the best decision rule. Secondly, usually the set of decision rules is replaced with some kind of real-valued (continuous) functions, because the minimization of the discrete empirical risk is most often infeasible. But as already mentioned, the minimizer of $\int_{\mathcal{Y}\times\mathcal{X}}(y - f(x))^2 \,\mathrm{dP}(y,x)$ among a set of real-valued functions $\mathcal{F}$ is the conditional expectation $f^*(x) = \mathbb{E}(Y = 1 \,|\, x)$, which in case of the binary random variable Y equals λ. However, if the minimization does anyway approximate the conditional probability function, then one could choose right at the beginning a more tailored approach towards the function approximation. So although argmin $\int(y - f(x))^2 \,\mathrm{dP}(y,x)$ and argmin $\int(\mathbb{E}(Y \,|\, x) - f(x))^2 \,\mathrm{dP}(x)$ have the same solutions, they both arise from quite different motivations. While the first formulation does target the problem in a statistical manner of explaining the observations, the second formulation is oriented to actually approximate a function.

Due to the inverse definition of the conditional probability function, the goal of function approximation associated to LUSI can be described with optimizing the following risk functional $\int(\mathrm{F}_1(x) - Af(x))^2 \,\mathrm{dP}(x)$. Instead of $F_1(x)$ only imprecise measurements $Y_i \,|\, x$ of it are available, with $\mathbb{E}(Y_i \,|\, x) = F_1(x)$, which arise from the empirical distribution estimator as shown. These circumstances do remind on the third kind of problems of dependency estimation introduced in the SL-theory, i.e. interpreting indirect experiments (sec. 2.2.1). However, different to the requirements stated there, the $Y_i \,|\, x = \hat{F}_1(x)$ are not independent in the current situation. Furthermore, the integral operator has to be estimated in case of LUSI too. Consequently, the problem can not be interpreted as an instance of the indirect experiments setting immediately. This is also why the empirically constructed optimization problem of LUSI is not an empirical risk as originally defined. This means, $\mathbb{E}(\,T_\gamma(\hat{A}f, \hat{F}_1)\,)$ does not represent the real risk functional, aimed to minimize, but rather does $T_0(\,\mathbb{E}(\hat{A}f, \hat{F}_1)\,)$, with T_γ being Tikhonov's regularization operator (eq. 3.2). The operator T_γ itself is not an estimate but a precise integration of two estimated functions. Also, the requirements of the regularization principle, as introduced in the previous chapter do differ from the application in LUSI. This is first because of the approximate operator. However, conditions for convergence can be found in [11]. Secondly, the regularizing operator is subject to randomness. A proof for convergence in probability of the approximate solution is given in [3, ch. 10].

3.2.2 (Statistical) Invariants

In the following, the concept of (statistical) invariants shall be investigated. In principle, the notion of the statistical invariants is an independent component of the LUSI method and could be combined with other models of estimating the conditional probability function too.

Why linear functionals?

In a general estimation setting, additional information about the function of interest $f^* \in \mathcal{F}$ could be given in different ways. One possibility is implicitly by knowing that for an operator (not necessarily linear or bounded) $A : \mathcal{F} \to \mathcal{W}$ the equation $A(f^*) = w$ holds. This information could then be incorporated by restricting the model space $\mathcal{F}$ to $\tilde{\mathcal{F}} = \mathcal{F} \cap A^{-1}w$. But different operator equations have different practicality. The relevant information is the set $A^{-1}w$. If for example neither A nor w is known, but this kind of equation shall still be incorporated, then it does matter how well A^{-1} can be approximated. In a statistical setting, this raises the question, whether/how well this set can be estimated. A canonical idea is to replace A and w by reasonable estimates and solve this approximate equation. However, this shifts the focus towards the existence of well behaved estimators for A and w. A second aspect of practicality is how well shaped the set $\tilde{\mathcal{F}}$ is, in the sense of how easily it can be optimized over it. This is because the implicitly defined function space might be quite difficult to actually be determined for any optimization solver. Concerning these two aspects a convenient situation is faced, if A is a linear functional over $L^2_{\mathrm{P}(x)}$, whose representation is called a predicate in the context of LUSI. In this case A as well w can easily be estimated through their representations as expectations as shown before. Furthermore, their empirical definitions deform to easily manageable linear side constraints for the optimization problem. Note, that in principle A could have been chosen independently of the unknown data distribution, such that $A(f)$ could explicitly be calculated for any $f \in \mathcal{F}$. However, then the problem would have been how to estimate $w = A(\lambda)$, since λ is unknown.

Another view on motivating the linear functionals is given in terms of weak convergence. In infinite dimensional, normed vector spaces $\mathcal{V}$, one differentiates between weak and strong convergence [4]. A sequence (v_n) does weakly converge to v, if $\forall v' \in \mathcal{V}' : \lim\limits_{n \to \infty} v'(v_n) = v'(v)$, at which the convergence is in the usual sense in the scalar space. Here $\mathcal{V}'$ denotes the dual space of $\mathcal{V}$ (app.A). The weak limit v does exist and is unique [16, p.117]. In case $\mathcal{V}$ is a Hilbert space then each linear functional is representable with an element of $\mathcal{V}$, such that the statement is equivalent to $\forall\, w \in \mathcal{V} : \langle v_n, w \rangle \to \langle v, w \rangle$. Comparing to the LUSI model, $\mathcal{V} = L^2_{\mathrm{P}(x)}(\mathcal{X})$, while $v = \lambda$ becomes the conditional probability function and $w = \psi$ is represented by any predicate. The associated linear functional is $v'_\psi(f) = \mathbb{E}(\psi(X)\, f(X))$. The sequence (v_n) corresponds to a sequence of estimated functions (f_l), for a sequence of samples of increasing size l. However, now also the functional v'_ψ is approximated denoted by $(v'_{\psi,l})$. As introduced, the estimators are $v'_{\psi,l}(f) = \frac{1}{l} \sum_{i=1}^{l} \psi(x_i)\, f(x_i)$, as well as $v'_{\psi,\lambda,l} := \frac{1}{l} \sum_{i=1}^{l} y_i\, \psi(x_i)$ for $v'_\psi(\lambda)$. If $v'_{\psi,l}(f_l) \to \lim\limits_{l \to \infty} v'_\psi(f_l)$ and $v'_{\psi,\lambda,l} \to v'_\psi(\lambda)$ converge as $l \to \infty$, then by satisfying the corresponding statistical invariants, it follows: $(\forall\, l : v'_{\psi,l}(f_l) = v'_{\psi,\lambda,l}) \implies v'_\psi(f_l) \to v'_\psi(\lambda)$ (applying the limit to both sides of the equation). Therefore, incorporating the respective statistical invariants could be interpreted as constructing a sequence of function estimates, which does (at least) weakly converge. This is a complement to the strong convergence, which is desired in the first place for function approximations. Of course

[4] Strong convergence means the convergence with respect to a metric.

the weak convergence is only an abstraction, since only a finite set of functionals can actually be considered in practice. Furthermore, the stochastic nature of the sequences should be noted, which is why any statements of convergence have to be appropriately adapted to random variables. Strong convergence does imply weak convergence [16, p.117]. This raises the question, what the gain in deploying weak convergence is. The reason is that the requirements for the estimators $v'_{\psi,\,l}$ and $v'_{\psi,\lambda,\,l}$ to converge appropriately are somewhat easier to satisfy. At first, it will be shown, that the required convergence of the empirical invariants is at least not harder to fulfill, than the uniform convergence of any empirical risk functional. Again it holds, that uniform convergence of $v'_{\psi,\,l}$ to v'_{ψ} is sufficient to guarantee the convergence of $v'_{\psi,\,l}(f_l) \rightarrow \lim_{l\rightarrow\infty} v'_{\psi}(f_l)$ [5]. Because $v'_{\psi,\,l}(f)$ is an empirical average of i.i.d observations, $v'_{\psi,\,l}$ conforms with the functional of empirical risk analysed in the SL-theory. Therefore, the previously derived conditions for convergence apply [6]. Especially if $\mathcal{F}$ has finite capacity, then the uniform convergence is given. Furthermore, the convergence of $v'_{\psi,\lambda,\,l}$ is basically matter of the laws of large numbers. But because the i.i.d.-property of the sample is anyway assumed in the context of the SL-theory, its convergence is implied. Hence the effect of the statistical invariants in terms of the convergence $v'_{\psi}(f_l) \rightarrow v'_{\psi}(\lambda)$ follows at least by the same requirements as for the empirical risk minimization. Another aspect of the convergence of the statistical invariants is the following. The estimator $v'_{\psi,l}$ is of a special form, as it only depends on the observations x_i. As mentioned previously, the relevant uncertainty in the analysis of empirical risk functionals is driven by $\mathrm{P}(y\,|\,x)$ (sec. 2.2.2). The assumptions made about $\mathrm{P}(x)$ are more of secondary relevance, also because one can take on a view of conditioning on fixed features x_i. Assuming $\mathrm{P}(x)$ is known, the value $\mathbb{E}(\psi(X)\,f(X))$ can either be calculated or is already determined by $v'_{\psi,\,l}$ itself. Thus, $v'_{\psi}(f)$ can directly be identified and it holds $v'_{\psi}(f_l) = v'_{\psi,\lambda,l}$. Therefore, the convergence $v'_{\psi}(f_l) \rightarrow v'_{\psi}(\lambda)$ depends only on the convergence of $v'_{\psi,\lambda,l}$ anymore. However, as just pointed out, because this is a simple empirical average the conditions of the respective laws of large numbers are sufficient. But these conditions are much less restrictive than the ones of uniform

[5] $|v'_{\psi,\,l}(f_l) - \lim_{l\rightarrow\infty} v'_{\psi}(f_l)| < |v'_{\psi,\,l}(f_l) - v'_{\psi}(f_l)| + |v'_{\psi}(f_l) - \lim_{l\rightarrow\infty} v'_{\psi}(f_l)|$

[6] Although the assumptions of proper boundedness (sec. 2.2.3) have to be met. However, for example in case of a continuous kernel function over a compact set $\mathcal{X}$ this is fulfilled.

convergence, as they are satisfied by available independent, identically distributed observations anyway [7].

Influence of invariants

Selecting m invariants means formulating m integral properties, which the approximate function should preserve. The associated value of the conditional probability function $\mathbb{E}(Y\,\psi(X))$ could be interpreted in the following way: $\psi(x)$ might describe some kind of property of the objects, whose expected value is quite characteristic for the pattern $Y = 1$. Thus, a good approximation of the conditional probability function should represent this behaviour, resulting in a restriction of the admissible function space. This is a fundamental difference to incorporating additional features in a model. While an increase of the dimension of x increases the capacity of the model, more invariants do decrease the capacity of the function space. As shown in the theory, a smaller capacity does in principle stand for a faster convergence rate of the empirical risks. However, shrinking the function space is not necessarily better, as it clearly depends what subset is selected. In case of precise knowledge of the invariants, they represent additional information about the desired function. Thus, in theory the function space would be shrunk in a meaningful direction, as the conditional probability function would still be contained. Therefore, the estimation would benefit from adding invariants to the model. In the statistical situation this is different. Due to the imprecision of the estimation, more statistical invariants do not necessarily shrink the function space for a given sample in a meaningful way. This is demonstrated by the following example: Consider l predicates, such that the following linear equation system results as constraints for the LUSI related optimization problem: $\Psi_{l\times l}\,(\mathbb{K}\,\alpha + c\,1_l) = \Psi_{l\times l}\,(y_1, ..., y_l)^\intercal$ (eq. 3.4). If Ψ has full rank, then the only admissible solutions are the functions, which fulfill $f(x_i) = y_i$. If $\mathbb{K}$ has full rank, then there is an appropriate vector α, such that this kind of function is representable in the chosen RKHS. Because this is the only admissible solution, the further optimization problem of LUSI already terminates. However, the function $f_l(x_i) = y_i$ just interpolates all the observations perfectly, but, as is known, does not converge to the conditional probability

[7] Although in case of conditioning, the $Y_i \,|\, x_i$ are not identically distributed.

function with increasing sample size. This example shows, that adding further statistical invariants does not necessarily improve the model. Hence, a trade-off regarding the selection of invariants results.

How to select invariants?

The main problem in the application of statistical invariants is the appropriate selection of finitely many predicates. This is especially driven by the trade-off between the idea of weak convergence (infinitely many invariants) and the empirical restrictions, as described before. There are no theoretical indicators, which give any criteria how to decide, what predicates the convergence of the empirical estimates to the conditional probability function will benefit more from. The selection should be made ideally problem specific. In the original paper, the selection of invariants is called the "intelligence driven" part of learning as complement to the data driven estimation. Proposing invariants is compared with a teacher-student interaction. Based on the teacher's domain specific insights he proposes meaningful characteristics. From a statistical point of view, the predicates should at least admit a good estimation. But this might be hard to tell, without assumptions about $\mathrm{P}(y, x)$. Especially, it has to hold that $\psi \in L^2_{\mathrm{P}(x)}$, in order for the respective expectation to even exist. The number and set of predicates do represent further hyperparameters in any model. Therefore, empirically a best subset of predicates could be determined based on generic procedures such as cross validation. Because the number of subsets is exponential in the total number of predicates, some kind of heuristic search might be useful. For this purpose the authors propose a stepwise selection, based on a modification, which will be presented below.

3.2.3 Substituting the $\mathbb{V}$-matrix

Due to the $\mathbb{V}$-matrix the pairwise products of the residuals influence the optimization problem's objective function (eq. 3.4). If these between-observations products were ignored, the objective would be equivalent to the common sum of squared residuals (except for the additive norm term), which results as empirical risk estimate of the risk functional $\int (y - f(x))^2 \, \mathrm{dP}(y, x)$. Hence, if the $\mathbb{V}$-matrix is replaced by the identity matrix in equation 3.4, i.e. $\mathbb{V} = \mathbb{I}$ [8], the corresponding model is the empirical regularized squared-loss minimization among a RKHS. This model is a special case of empirical risk minimization and constitutes a classical approach to the pattern recognition problem. This relation is useful for evaluating the LUSI model, because it admits a comparable approach in a canonical way. Therefore, the immediate effect of deploying the $\mathbb{V}$-matrix, as result of the inverse-problem based approach of estimating the conditional probability function, can be assessed. Although both methods emerge from different backgrounds their empirical loss functionals are quite similar. This relation might indicate, that both methods provide similar approximation as well as classification accuracy.

3.3 Modifications

In the following, two modifications of the LUSI method, introduced in the original paper [11], are presented and further analysed. They arise quite intuitively and are important for the practical utilization and final evaluation.

3.3.1 Relaxing statistical invariants

As already pointed out, the statistical invariants lack both certainty of the right-hand side and left-hand side of the equations. To counteract the estimation variance a simple solution can be applied. Instead of demanding a function satisfying the statistical invariants exactly, a (small) deviation is allowed. Concretely, the respective side constraints become inequalities: $\frac{1}{l}\,|\sum_{i=1}^{l} \psi_j(x_i)\, f(x_i) - \sum_{i=1}^{l} y_i\, \psi_j(x_i)| \leqslant \delta$. In the following the absolute difference between right-hand side and left-hand side of a statistical invariant is called the invariant's discrepancy/deviation. The deviation δ has to be specified in advance, either as absolute value or as relative deviation: $\delta_{abs} = \delta_{rel}\,|\frac{1}{l}\,\sum_{i=1}^{l} y_i\, \psi_j(x_i)|$. By introducing the inequalities the probability of a function belonging to the admissible set of the optimization problem becomes actually positive. When allowing deviations, the delta value δ becomes a further hyperparameter to control the estimation. Immediately the question of how to set them appropriately arises. For this purpose some as-

[8] $\mathbb{V}_{i,j} = 1$ if $i = j$, else it is 0

pects shall be mentioned. At first, a simple theoretical bound on the delta's value is derived. Therefore, notice that $|\sum_{i=1}^{l}\psi(x_i)\, f(x_i) - \sum_{i=1}^{l} y_i\, \psi(x_i)| = |\sum_{i=1}^{l}\psi(x_i)\,(f(x_i) - y_i)| \leqslant \sqrt{\sum_{i=1}^{l}\psi(x_i)^2}\, \sqrt{\sum_{i=1}^{l}(f(x_i) - y_i)^2}$, which follows by Cauchy-Schwartz' inequality. Since $f(x_i)$ is bounded between $[0,1]$ and y_i is from $\{0,1\}$, it holds: $(f(x_i) - y_i)^2 \leqslant 1 \implies \sqrt{\sum_{i=1}^{l}(f(x_i) - y_i)^2} \leqslant \sqrt{l}$. Together it results:

$$|\tfrac{1}{l}\sum_{i=1}^{l}\psi(x_i)\, f(x_i) - \tfrac{1}{l}\sum_{i=1}^{l} y_i\, \psi(x_i)| \leqslant \sqrt{\tfrac{1}{l}\sum_{i=1}^{l}\psi(x_i)^2} \tag{3.5}$$

Consequently for the relative deviation follows: $\delta_{rel} \leqslant \sqrt{\frac{\frac{1}{l}\sum_{i=1}^{l}\psi(x_i)^2}{\frac{1}{l}\sum_{i=1}^{l}(y_i\,\psi(x_i))^2}}$. These values can be consulted, to obtain a precise bound, up to what any value for δ has to be searched for. Furthermore, the inequality by Cauchy-Schwartz reveals, that the function, which interpolates all observations perfectly does always satisfy the statistical invariants and is therefore an admissible solution for the LUSI optimization problem. This can also be verified directly by taking a look into the formulated optimization problem (eq. 3.4). A sufficient condition, that this kind of function is element of the function space is, that the matrix $\mathbb{K}$ has full rank. The delta value could also be specified on the basis of some probability informed confidence. For this purpose one could consider the following inequality, with $i_{\psi,f} := \mathbb{E}(f(X)\,\psi(X))$:

$$\begin{aligned} &\tfrac{1}{l}\,|\sum_{i=1}^{l}\psi(x_i)\, f(x_i) - \sum_{i=1}^{l} y_i\, \psi(x_i)| \\ &= |\tfrac{1}{l}\sum_{i=1}^{l}\psi(x_i)\, f(x_i) - i_{\psi,f} + i_{\psi,f} - i_{\psi,\lambda} + i_{\psi,\lambda} - \tfrac{1}{l}\sum_{i=1}^{l} y_i\, \psi(x_i)| \\ &\leqslant |\tfrac{1}{l}\sum_{i=1}^{l}\psi(x_i)\, f(x_i) - i_{\psi,f}| + |i_{\psi,f} - i_{\psi,\lambda}| + |i_{\psi,\lambda} - \tfrac{1}{l}\sum_{i=1}^{l} y_i\, \psi(x_i)| \end{aligned} \tag{3.6}$$

The first and the third term describe the variation of the empirical averages around the respective expectation values. The third term is constant across all functions f. The second term is the offset between the linear functional's value at λ and the corresponding value for f. If it is possible to derive meaningful bounds on the ranges (moments) of these terms, general intervals could be derived by means of

Hoeffding's or Chebychev's inequalities [9]. Another approach to bound the delta value is given by the fact, that in order to enable any difference by incorporating the invariants into the model, the allowed discrepancy must be smaller than the achieved deviation of the optimal solution, when ignoring the statistical invariants at all. Therefore, one could solve the optimization problem twice. First, ignoring the invariants at all. Afterwards, for the received solution, the discrepancy of the statistical invariants is calculated. This value can then be used to bound the allowed deviation in the second run, when actually considering the invariants. In practice some generic approach such as cross-validation might be preferred.

An extension is to allow individual deviations for each predicate δ_ψ. By manipulating these deltas the respective invariant's influence on the optimization problem can be steered. This enables some kind of selection method among a set of predicates, since allowing big deviations is equivalent to ignoring the respective invariant. However, the drawback of this approach is the higher number of parameters in the model. Based on the discrepancies, the authors proposed a heuristic selection among a pre-specified set of predicates, in terms of a stepwise selection process. At first a general threshold is defined. For a given function all invariants, whose discrepancy is smaller than the specified threshold are considered as satisfied. Beginning with no invariants at all, the optimization problem is solved. For the received estimated function, the invariant with the highest discrepancy bigger than the threshold is added to the optimization problem. Subsequently, a new function is estimated. This process continuous, until all invariants are satisfied with at least the specified threshold accuracy.

3.3.2 Reweighted $\mathbb{V}$-matrix

Although the LUSI method aims to estimate the conditional probability function, it is still placed in the context of pattern recognition. Therefore, the authors proposed a modification to shift the focus back towards the classification accuracy. The modification bases on the observation, that for a good classification the prob-

[9] For example $f(x_i)$ as well λ lie between 0 and 1. Also, for a continuous predicate ψ over a compact interval $\mathcal{X}$ a bound exists too.

ability should precisely be estimated especially at points x with $\lambda(x) = 0.5$. This is because 0.5 is the switchover point for the associated decision rule. For a true probability of 0.8 it is of not much relevance whether the estimate is 0.6 or 0.9 as long as it is bigger than 0.5. To put more weight onto these points, the $\mathbb{V}$-matrix is modified in the following way. The basic idea is to give observations y_i, whose conditional variance is high a bigger weight. This is because, the variance is biggest if the conditional probability is 0.5, but zero if it is 0 or 1. For this purpose, the $\mathbb{V}$-matrix is adjusted to: $\tilde{\mathbb{V}}_{i,j} = \mathbb{V}_{i,j}\, o(\sigma(x_i))\, o(\sigma(x_j))$. Where o is a monotonically increasing function and $\sigma(x)$ is the standard deviation of $Y \mid x$. o is usually set to the identity function. By this adaption a new optimization problem is defined, putting more weight on observations with conditional probability similar to 0.5 and less weight for more extreme probabilities. Typically, the standard deviation has to be estimated. Therefore, the complete procedure consists in two stages. At first the conditional probability function is estimated based on the original presented LUSI model. Then the standard deviation is estimated according to its definition for binary random variables: $\sigma(x) = \sqrt{\lambda(x)\,(1 - \lambda(x))}$, replacing λ with the estimated function. According to the adjusted optimization problem a new function is estimated, based on which a decision rule is derived.

4 Practical evaluation

Having introduced the theoretical backgrounds of statistical learning as well as the LUSI model of estimating the conditional probability function, it follows the presentation of the practical part of this thesis. The purpose is to evaluate the potential of the new method, together with providing an implementation for further independent use. In the original paper a simulation study was conducted, to analyse the advantages of the newly introduced method against various variations of it. The objective of the presented study is to verify the original results, by reconstructing the tests as far as possible. Therefore, analysing how reliable the claimed improvements really are. As will be pointed out soon, the original study was not conducted very profoundly and has some critical weaknesses. Especially, some important information of the context of testing were not documented. This is why, in this thesis the tests were considerably extended to receive more expressive and consistent statements about the method's accuracy. According to the two purposes behind the LUSI method, two different criteria were analysed in the paper. These are on the one hand the estimation quality of the conditional probability function and on the other hand the accuracy of classification. The LUSI method is essentially characterized by two components, whose effect shall be assessed. These are the $\mathbb{V}$-matrix, as consequence of the inverse-problem based function approximation by Tikhonov and the statistical invariants.

4.1 Configurations of testing

In this section the configurations of the undertaken simulation study are presented. It is contrasted what conditions were specified by the original paper and what the finally chosen settings for this thesis were. Furthermore, the reasoning behind the specification of the various parameters is provided, whenever the paper did not give any precise information. The general procedure of the conducted study in this thesis is summarized as follows. The study was simulation-based, which means that the data generating distribution is known and accessible. Besides the LUSI method other methods were defined for comparison. For each such model class, the regularization parameter γ and the delta value δ in the invariants' constraints define the concrete optimization problem (eq. 3.4). For both hyperparameters a set of values was preassigned [1]. Each combination of these values determined a specific model, which was evaluated. The goal lied in assessing the expected value of two different evaluation-metrics for each of the models for different sample sizes. To estimate these expectations, the calculations were repeated for multiple independent samples, over which an average was finally determined. The table 4.1 shows the relevant parameters of the conducted study, together with the values according to the original study and the final applied ones. In the following, the various parameters are explained in more detail. For the remaining part, the terminology shall uniquely be defined. For this purpose, a model is meant to be an estimator, mapping from the space of samples into $\mathcal{F}$. While a model class is in general an arbitrary set of models. This is for example canonically defined, by specifying a range of values for some hyperparameters, here essentially γ and δ.

Evaluation-metrics

Each estimated function was evaluated with respect to two risk functionals. These were the L^2 metric to the conditional probability function and the probability of false classification (error probability) of the derived decision rule. In the following the evaluation of any of these risk functionals at a specific function is also called the function's score. Because the data distribution was accessible, the real risk values according to the following formulas could be calculated: $l_2(f) := \sqrt{\int_{\mathcal{X}}(f(x) - \lambda(x))^2 \,\mathrm{d}x}$, $err(r) := \int_{\mathcal{X}}(y - r(x))^2 \,\mathrm{dP}(y, x)$, with $f \in \mathcal{F}$ and r being the associated decision rule (sec. 3.1). Note, that the function distance is in L^2 and not in $L^2_{\mathrm{P}(x)}$, as was specified by the paper. However, no information was given, whether these integrals were estimated based on the sample data or by actual (numerical) integration. Therefore, in this study the actual numerical

[1] In the following the values of γ and δ are referred to as hyperparameters.

Parameter	Value according to [11]	Final setting
Evaluation metrics	L^2 distance to λ False classification probability (error probability)	
Sample size	$l = 48, 96, 192$	$l = 25, 50, 100, 200$
Number of samples	No information, presumbly 1	200
Data distribution	-λ: No information, but coarse visualization is given -P$(Y = 1) = {}^{1}\!/_{3}$ -$\mathcal{X} = [0, 1]$, one dimensional -P(x): No information, only rug plot of a sample	-λ: polynomial interpolation based on points derived from visualization -P$(Y = 1) = 0.33338$ -$\mathcal{X} = [0, 1]$, one dimensional -P(x): Beta(1.98, 2.16), chosen to achieve the marginal probability of Y
Model classes	SVM: $\mathbb{V} = \mathbb{I}$, no invariants vSVM: no invariants SVM_I2: $\mathbb{V} = \mathbb{I}$, two invariants vSVM_I2: two invariants mSVM_I2: $\mathbb{V} = \mathbb{I}$, reweighting the identity matrix, two invariants mvSVM_I2: modification of reweighting, two invariants	SVM, vSVM, SVM_I2, vSVM_I2, mSVM_I2, mvSVM_I2, SVM_I4, vSVM_I4
Predicates	$\psi_1(x) = 1$, $\psi_2(x) = x$, no selection	$\psi_1(x) = 1$, $\psi_2(x) = x$, $\psi_3(x) = x^2$, $\psi_4(x) = x^3$, no selection
Kernel function	INK Kernel, degree 1, bandwidth 1	
Gamma γ	No information	$\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$
Deltas δ	No information, presumbly none	for all predicates the same value, specified as relative difference $\{0, 0.05, 0.1, 0.2, 0.4, 0.8\}$

Table 4.1: Configurations of the simulation study.

integration over $\mathcal{X}$ for both types was considered [2]. To do so, the total interpolated function is required, instead of only the estimated function's value at the observations x_i. According to the introduced representer theorem, the estimated functions of the RKHS are defined by $f(x) = \sum_{i=1}^{l} \hat{\alpha}_i \, K(x, x_i) + \hat{c}$. $(\hat{\alpha}, \hat{c})$ are the minimizers of the respective optimization problem. One should be aware of the difference between the risk of an estimated function of the function space $\mathcal{F}$, denoted here as score and the risk of a model (estimator). For a model the measure of interest is usually the expected value of the scores over the space of possible samples.

Sample size

In the paper the models were evaluated on three different sample sizes. To gain further insights, a fourth sample size of $l = 25$ was considered at this point, to analyse the effects of very little information.

Number of samples

The authors of the original paper did not make any clear statements, how their presented values were obtained. It is indicated, that the experiments were only performed on one sample (for each sample size). This is a tremendous weakness, because any kind of uncertainty quantification was left out. To overcome this problem, the here presented study is based on multiple repetitions. Concretely 200 samples were chosen for each sample size. This allowed to estimate confidence intervals, providing a much more consistent insight. The choice of the concrete number of samples was influenced by the trade-off between accuracy and computational effort. To make an informed decision a representative model was chosen and the complete evaluation performed on N samples. The uncertainty of the received average scores was assessed by the respective variation coefficient as well as the relative confidence interval length. Based on these indicators a number N was sought, such that these quantities fall below a minimum level. On top of this, the number of 200 was set to exhaust the computational power.

[2] The calculated scores for a given function are the real expectations, up to some approximation error, due to the numerical integration. It is not an empirical estimate.

Data distribution

Another difficulty in reconstructing the original study arose from the incomplete documentation of the applied data distribution. The domain $\mathcal{X}$ was given as $[0, 1] \subseteq \mathbb{R}$. Furthermore, it was specified, that the marginal probability $\mathrm{P}(Y = 1)$ was one third. The authors chose this, to analyse the robustness in scenarios of class imbalance. However, the marginal distribution $\mathrm{P}(x)$ and the actual conditional probability function λ were not published. To approximate the conditional probability function, a cubic spline interpolation was carried out based on 10 points, which were derived from (rather rough) visualizations occurring in the paper. The original function and the constructed approximation can be seen in figure B.2. This approximation should be good enough to not distort the final statements. Due to $\mathcal{X} = [0, 1]$ a beta distribution was chosen for the x-variable. The concrete density was specified, to satisfy $\mathrm{P}(Y = 1) = \int_{\mathcal{X}} \lambda \, \mathrm{p}(x) \, \mathrm{d}x = \frac{1}{3}$. The density distribution is plotted in figure B.1. In the paper a rug plot over the x-axis of the used sample is visible. The main amount of observations is located in the middle, while only few observations are at the extremes, i.e. < 0.1, > 0.9. This gives at least some indication, that an uniform distribution is not plausible. Based on the rug-plot, one may carefully deduct, that $\mathrm{P}(X > 0.9) \approx 0.02$, which is the same as for the chosen beta distribution.

Model classes

The LUSI method is characterized by the $\mathbb{V}$-matrix, weighting the pairwise residual products and the m invariants defining linear side constraints. This model class is denoted by "vSVM_Im" [3]. In the paper two invariants were used (m=2), which are explained below. To compare the method appropriately, the authors considered four alternatives, which differ in the use of the $\mathbb{V}$-matrix or its substitution with the identity matrix and the inclusion or ignorance of the invariants. These alternatives are denoted by either the presence or absence of "v" or "I2" respectively. They are appropriate for comparison, as these models differ in only single components, which characterize the LUSI approach. Thus, the immediate effect of the inverse-problem based function approximation or of the invariants can be identified. As

[3] The notations should not be confused with the common SVM classifier, although there are certain similarities. Instead, these were overtaken from the original paper.

explained before, the models substituting the $\mathbb{V}$-matrix (any without "v"), base upon the estimation of the conditional probability function according to the empirical risk minimization of the regularized squared-loss risk functional (sec. 3.2.3). The authors did also consider the modified version of reweighting the $\mathbb{V}$-matrix, which is denoted by "m(v)SVM_I2". As mentioned before, this method was introduced with focus on the classification (sec. 3.3). Note, that this adjustment is independent of the approach how to estimate λ. This is why, "mSVM_I2" is the model class of regularized squared-loss minimization, at which the reweighting is applied to the identity matrix. To extend the original analysis this study considers two further methods, which arise by adding two more invariants, i.e. "SVM_I4", "vSVM_I4". This shall provide an improved intuition, whether the invariants are indeed enhancing the ground methods, but also whether a sub selection among the predicates can be useful.

Predicates

As mentioned, the creation of predicates is in its core a knowledge-driven part and should be ideally problem specific (sec. 3.2.2). Because it is not known, how to decide what predicates are advantageous, the authors proposed two generic moments-based predicates: $\psi_1(x) = 1$, $\psi_2(x) = x$. The predicate ψ_1 implies the invariant: $\mathbb{E}(1\, f(X)) = \int_{\mathcal{X}} f(x)\,\mathrm{dP}(x) = \mathrm{P}_{\lambda=f}(Y = 1) \overset{!}{=} \mathrm{P}_{\lambda}(Y = 1)$ [4]. This means for the empirical version a hypothesis f is sought, such that the estimate of the marginal probability of class one considering f as the conditional probability function is equivalent to the mean observed in the data. The second invariant develops to: $\mathbb{E}(X\, f(X)) = \int_{\mathcal{X}} x\, f(x)\,\mathrm{dP}(x) = \mathbb{E}_{\lambda=f}(YX) \overset{!}{=} \mathbb{E}_{\lambda}(YX)$. Together with the first invariant, this relates to finding a hypothesis, such that the conditional expectation $\mathbb{E}(X \mid y = 1)$ is maintained. Following this idea of constructing predicates two further moments were considered: $\psi_3(x) = x^2$, $\psi_4(x) = x^3$. Therefore, demanding to preserve the conditional variance and skewness.

Kernel function

The function space $\mathcal{F} = H_K$ is uniquely defined by its kernel. The authors applied the following kernel: $K_1(x_1, x_2) = \frac{1}{3}\,(\min(x_1, x_2))^3 + \frac{1}{2}\,(\min(x_1, x_2))^2\,|x_1 - x_2|$. This

[4] The second equation results, because f is a conditional probability function itself.

kernel is a special case of the INK-kernel with degree 1. The INK-kernel is built up by splines of degree $r \in \mathbb{N}$ with infinite knots [17]. Usually it is defined without any bandwidth (scaling) parameter. Therefore, motivated by other kernels, such as the well-known Gaussian-kernel, one might think to reshape it by introducing a parameter b, such that: $\tilde{K}_{\text{INK}}(x_1, x_2) := K_{\text{INK}}(b\,x_1, b\,x_2)$. As can easily be verified, this scaling parameter b is due to the polynomials equivalent to using a multiplicative coefficient b^{2r+1}, thus: $K_{\text{INK}}(b\,x_1, b\,x_2) = b^{2r+1}\,K_{\text{INK}}(x_1, x_2)$ [5]. Consider the defined LUSI optimization problem $\text{OP}_{\text{LUSI}}(\gamma, \delta, b)$, with the parameters γ, δ and the bandwidth b. Then, due to the derived property of the INK-kernel regarding the bandwidth, it holds: $\text{OP}_{\text{LUSI}}(\gamma, \delta, b) \equiv \text{OP}_{\text{LUSI}}(\gamma\,\frac{1}{b^{2r+1}}, \delta\,\frac{1}{b^{2r+1}}, 1)$. This can easily be verified, by executing the respective substitutions in equation 3.4. This relation shows, that one can indeed ignore the scaling parameter and fix it to 1 and control the optimization problem via γ and δ.

Regularization parameter γ

The regularization parameter determines the weight of the function norm onto the objective in the optimization problem. Therefore, it influences the complexity of the estimated function, which is mirrored in its smoothness. If γ is set to 0 the solution of the optimization problem will be a function, which perfectly mirrors the observations. Note, that this function is also always an admissible function of the invariants, as shown before (sec. 3.3). Again, the authors did not specify, what γ parameter they applied. In this study a set of values was predefined and each corresponding model evaluated. The tested values were determined after assessing on individual samples the influence of the γ parameter. The prior investigation showed, that values bigger than 10 did not affect the estimated function anymore, as the function with minimal norm was always chosen.

Deltas δ

The delta value defines the allowed discrepancy in the linear side constraints in the optimization problem defined by the statistical invariants. Again, no information

[5] $x = (x_1, ..., x_d)$, $y = (y_1, ..., y_d) \in \mathbb{R}^d$

$$K_{\text{INK},r}(x, y) := \prod_{i=1}^{d} K_r(x_i, y_i) \quad K_r(x_i, y_i) := \sum_{k=0}^{r} \frac{C_r^k}{2r-k+1}\,[\min(x_i, y_i)]^{2d-k+1}\,|x_i - y_i|^k \text{ [17]}$$

about its specification was given by the authors. Instead, it can be assumed, that strict equality was required, although differently recommended by the authors themselves. However, in this thesis different deviations were tested. To reduce the complexity the same relative absolute deviation was used for all invariants. The higher the allowed discrepancy, the less is the effect of deploying the invariants at all. An upper bound of the δ values to be considered, was determined, by estimating the 0.75 quantile of the maximum relative absolute deviation of all statistical invariants for the defined conditional probability function, based on 200 samples [6]. This way, the conditional probability function should remain an admissible function in a significant amount of times. Furthermore, as for the γ parameter the effect of the δ value was examined on individual samples and models. This way, the threshold upon which the invariants could effectively be ignored, because the allowed deviation represents no constraint anymore, was tried to be determined.

4.2 Implementation

In the following, some information about the implementation as well as required technical adaptations are explained.

Module structure

The guiding idea of the implementation was to enable both the separate use of the LUSI method of its evaluation and the possibility to conveniently modify or extend the simulation study. Therefore, the LUSI method and its evaluation are implemented with little interdependency. The simulation infrastructure is hold generic as far as possible, such that all settings displayed in table 4.1 can easily be varied. The implementation is realized in R. For the (quadratically) optimization problem the Gurobi framework is deployed. The implementation is structured into four main modules. Their main functionality is summarized in the table 4.2.

[6] These were different, than the ones used for the final evaluation.

File	Functionality
LUSI.R	The file contains the implementation of the LUSI model and the presented various modifications. Both kernels, i.e. the Gaussian- and the INK-kernel of variable degree are implemented. Furthermore, the calculation of the L^2 metric and the accuracy of classification based on one-dimensional numerical integration as well as the construction of the complete function of the RKHS for given coefficients are realized. The main function is called *lusi()*. Its parameters allow to create each modification presented in the model-classes for the evaluation. Additionally, aspects such as the customizable stabilization of the kernel matrix and the proposed algorithm for predicate selection of the original paper are implemented. Furthermore, it is possible to define individual deviations for each invariant based on an absolute or relative specification. It can be specified based on which formulation the optimization problem shall be solved together with various parameters to adjust the optimization solver.
simulation_study.R	This file does implement the necessary functionality for the simulation. The central function is called *hyperpar_testing()*. Among others it takes a list of four lists. Each sublist contains the testable values for one of these parameters: regularization parameter γ, delta values $\delta_1, ..., \delta_m$ for the invariants, bandwidth of kernel, quality threshold for predicate selection. The remaining attributes specify a set of samples over which to iterate and the concrete model class which shall be used. The function outputs as many matrices as score functions defined. Each matrix contains as many rows as samples and as many columns as hyperparameter combinations. Each cell contains the score value of the estimated function on this sample. Furthermore, the file contains the definition of the data distribution, the predicates and the evaluation-metrics, which can be applied to a single function using the implementation of the *LUSI.R* file.
simulation_models.R	This file serves the definition of the simulation setup, which shall be executed. The *hyperpar_testing()* function is specified and called. The obtained results of the simulation are persistently stored on the hard disc.
simulation_analysis.R	Here the necessary functionality for evaluating the obtained score matrices is implemented. The file's structure is rather complex. It contains functionality for calculating different confidence intervals for averages, as well as for quantiles. The main function *model_evaluation()* calls the function *model_analysis()* and summarizes the distribution of the score values of the best model for each model class. The function *plots_and_pairwise_eval()* serves the pairwise comparison of the best models. It can be determined, whether the evaluation should be based on averages or quantiles. Besides the numerical statistics also corresponding plots are created.
combine_experiments.R	Auxiliary file, to combine results of different runs for the same model class.
refitting_best_models.R	Auxiliary file, to refit the best models on new samples.

Table 4.2: Summary of the implementation. Presented are the file names together with an overview of the main functionalities.

Problem transformation

The interface of Gurobi requires the optimization problem in a canonical formulation [18]: $\min\limits_{v} v^\intercal \mathrm{Q}\, v + c^\intercal v \quad \text{s.t.}\ \mathrm{M}v\ <>= \ \beta$ where Q, M are the matrices defining the quadratic objective and the linear side constraints respectively, while c, β are vectors defining the linear component of the objective and the thresholds of the linear constraints. The operational relations can be specified for each constraint individually. To satisfy the required form, the LUSI optimization problem as introduced (eq. 3.4), has to be transformed. Working in the relative delta value δ (sec. 3.3), the following optimization problem results:

$$\tilde{\alpha} := (\alpha^\intercal, c)^\intercal_{(l+1)\times 1} \ \mathbb{K}_1 := \begin{pmatrix} \mathbb{K} & 1_l \end{pmatrix}_{l\times(l+1)} \ \mathbb{K}_2 := \begin{pmatrix} \mathbb{K} & 0_l \\ 0_l^\intercal & 0 \end{pmatrix}_{(l+1)\times(l+1)}$$

$$\begin{aligned} \min_{\alpha,\, c}\ & (\mathbb{K}\,\alpha + c\,1_l)^\intercal\, \mathbb{V}\, (\mathbb{K}\,\alpha + c\,1_l) - 2\,(\mathbb{K}\,\alpha + c\,1_l)^\intercal\, \mathbb{V}\, y + \gamma\, \alpha^\intercal \mathbb{K}\, \alpha \\ &= (\mathbb{K}_1\, \tilde{\alpha})^\intercal\, \mathbb{V}\, (\mathbb{K}_1\, \tilde{\alpha}) - 2(\mathbb{K}_1\, \tilde{\alpha})^\intercal\, \mathbb{V}\, y + \gamma\, \tilde{\alpha}^\intercal\, \mathbb{K}_2\, \tilde{\alpha} \\ &= \tilde{\boldsymbol{\alpha}}^\intercal\, (\mathbb{K}_1^\intercal\, \mathbb{V}\, \mathbb{K}_1 + \gamma\, \mathbb{K}_2)\, \tilde{\boldsymbol{\alpha}} + (-2\, y^\intercal\, \mathbb{V}\, \mathbb{K}_1)\, \tilde{\boldsymbol{\alpha}} \\ \text{s.t.} & \left(\begin{pmatrix} \Psi_{m\times l} \\ \Psi_{m\times l} \\ \mathbb{I}_{l\times l} \\ \mathbb{I}_{l\times l} \end{pmatrix} \mathbb{K}_1 \right) \tilde{\boldsymbol{\alpha}} \begin{pmatrix} \leqslant \\ \geqslant \\ \leqslant \\ \geqslant \end{pmatrix} \begin{pmatrix} \Psi\, y + \delta\,(\Psi\, y) \\ \Psi\, y - \delta\,(\Psi\, y) \\ 1_l \\ 0_l \end{pmatrix} \end{aligned} \tag{4.1}$$

Numerical aspects

First experiments revealed some numerical instability. The main causation was the kernel matrix. Depending on its rank or condition respectively the Gurobi solver did sometimes not converge in an optimal state, although an optimal admissible solution should exist. Further investigations showed, that this is especially a problem when using the Gaussian-kernel. Then, the rank heavily depends on the bandwidth. For the finally applied INK-kernel, the matrix showed almost always full rank, such that the problem was not as much severe. However, to counteract this behaviour the following actions were implemented. As explained in theory, the kernel matrix is symmetric as well as positive semi-definite by definition. Therefore, a small constant of 10^{-6} is added to the diagonal of the kernel matrix to make it regular. The implementation allows to either force this stabilization independently of the actual rank or, as finally used, to add it only if the rank is incomplete. The constant should be small enough to not distort the results with respect to the following equation: $\mathbb{K}\,(\mathbb{K}+q\,\mathbb{I})^{-1} \approx \mathbb{I}$, $q \in \mathbb{R}^{+}$. This is especially relevant for the construction of the total function based on the coefficient estimates. The coefficients $\tilde{\alpha}$ are obtained with respect to the specification of the optimization problem using $(\mathbb{K} + q\,\mathbb{I})$. However, the corresponding total function, which is finally evaluated, is defined by means of the partial functions $K_{x_i}(x) = K(x, x_i)$ (sec. 2.3.6). Hence, for the evaluation at x other than x_i the offset q cannot be incorporated (as it is meaningless) and has to be ignored for the interpolation. The official Gurobi documentation recommends some guidelines, how to deal with numerical problems [19]. According to these, the following parameter configurations were used: *NumericalFocus*=2, *BarHomogenous*=1. Furthermore, especially the ranges and the magnitudes of the matrices specifying the optimization problem are of importance. According to the documentation the most effective action is to reformulate or simplify the problem, to avoid unnecessary calculations or to receive bounded variables. For this purpose, the optimization problem is transformed, by substituting the estimated function's values at the observations x_i by an individual vector:

$$
\begin{aligned}
&\mathbb{K}_q := \mathbb{K} + q\,\mathbb{I} \quad p := \mathbb{K}_q\,\alpha + c\,1_l \quad \tilde{p} := (p^\intercal, c)^\intercal_{(l+1)\times 1} \quad \tilde{\mathbb{I}} := \begin{pmatrix} \mathbb{I} & \text{-}1_l \end{pmatrix}_{l\times(l+1)} \\
\min_{p,\,c}\; &(\mathbb{K}_q\,\alpha + c\,1_l)^\intercal\,\mathbb{V}\,(\mathbb{K}_q\,\alpha + c\,1_l) - 2\,(\mathbb{K}_q\,\alpha + c\,1_l)^\intercal\,\mathbb{V}\,y + \gamma\,\alpha^\intercal\,\mathbb{K}_q\,\alpha \\
&= p^\intercal\,\mathbb{V}\,p\; - 2\,p^\intercal\,\mathbb{V}\,y + \gamma\,(p - c\,1_l)^\intercal\,\mathbb{K}_q^{-1}\,(p - c\,1_l) \\
&= \tilde{\boldsymbol{p}}^\intercal\,(\begin{pmatrix} \mathbb{V} & 0_l \\ 0_l^\intercal & 0 \end{pmatrix} + \gamma\,\tilde{\mathbb{I}}^\intercal\,\mathbb{K}_q^{-1}\tilde{\mathbb{I}})\,\tilde{\boldsymbol{p}} + \begin{pmatrix} -2\,y^\intercal\,\mathbb{V} & 0 \end{pmatrix}_{1\times(l+1)}\,\tilde{\boldsymbol{p}} \\
\text{s.t.}\; &\begin{pmatrix} \Psi_{m\times l} & 0_m \\ \Psi_{m\times l} & 0_m \\ \mathbb{I}_{l\times l} & 0_l \\ \mathbb{I}_{l\times l} & 0_l \end{pmatrix} \tilde{\boldsymbol{p}} \begin{pmatrix} \leqslant \\ \geqslant \\ \leqslant \\ \geqslant \end{pmatrix} \begin{pmatrix} \Psi\,y + \delta\,(\Psi\,y) \\ \Psi\,y - \delta\,(\Psi\,y) \\ 1_l \\ 0_l \end{pmatrix}
\end{aligned}
\tag{4.2}
$$

This substitution brings some advantages. At first, it reduces the number of matrix multiplications, because it cancels out the matrix $\mathbb{K}_1$ (eq. 4.1). The second point is, that the variables p, representing the estimated probabilities at the observations x_i, are naturally bounded. This property is favourable for the solver routine, than the unbounded variables α before. This formulation requires the kernel matrix $\mathbb{K}$ to be invertible. However, due to the implemented stabilization $\mathbb{K}_q$ this is always guaranteed. After having estimated the function's values p, the corresponding coefficients of the actual function are derived, by calculating $\mathbb{K}_q^{-1}\,(p - c\,1_l)$. Furthermore, in the optimization problem the boundedness of the estimated function's values can only be controlled at the observations. This is why it is possible, that some values of the interpolated function are below zero or bigger than one respectively. In this case, the implementation proceeds by truncating the values at zero or one.

Running time

Based on the presented ranges (tab. 4.1) for the hyperparameters γ and δ, for each model class, which considers the invariants, up to 36 models result. For the other cases it is 6 models. All models were evaluated on 200 samples for four sample sizes and two scores were calculated based on numerical integration for each estimated function. This led to some considerably amount of computational effort. However, the whole process can be executed for each sample independently. This fact was exploited, in order to reduce the duration, by implementing a parallel computation across the samples. The total time required for running all experiments and associated evaluations was still about 60 hours on an Intel(R) Core(TM) i5-8265U CPU at 1.60GHz and 64-Bit system.

4.3 Methodology of evaluation

All model classes were tested on the same set of 200 samples. For each model, implied by a combination of the hyperparameters γ and δ, the corresponding optimization problem was solved. Subsequently, the L^2 distance of the estimated function to the conditional probability function as well as the error probability of classification for the associated decision rule were calculated. The result of these simulations can be imagined as two matrices for each model class, one for each type of the presented evaluation-metrics. A matrix has as many columns as tested hyperparameter combinations and as many rows as used samples. In the following, the methodology of analysing these received score matrices will be explained. In general, the procedure is about to identify for each model class a best model for each type of the evaluation-metrics (tab. 4.1). Subsequently, these best models are compared across the model classes. Selecting a best model means facing a decision problem. The evaluation bases on two different approaches of how to define the performance of a model(-class).

Priori evaluation [7]

The most intuitive way is to evaluate the models based on the expected score value, e.g. $\mathbb{E}(l_2(\hat{\lambda}_{\gamma,\delta}((Y_i, X_i)^{(l)})))$ (sec. 4.1), at which $\hat{\lambda}_{\gamma,\delta}$ depicts the model as an estimator mapping from the space of samples into $\mathcal{F}$. Consequently, a model class is represented by the risk value $\min\limits_{\gamma_j,\delta_j} \mathbb{E}(l_2(\hat{\lambda}_{\gamma_j,\delta_j}((Y_i, X_i)^{(l)})))$, at which γ_j, δ_j are the considered values of the respective models in the model class [8]. This approach might be interpreted as an "a-priori" selection, because the evaluation of a model is based on the marginal expected performance, which considers the variation in the estimation. Thus, if in practice there was enough information to calculate these expectations, the best model would be determined independently (a-priori) of the given sample. To estimate the risk of a model, the empirical average of the respective 200 score values is calculated. According to these averages, for each model class a best model is identified regarding the lowest L^2 distance and a best model regarding the lowest error probability of classification.

Posteriori evaluation

Whenever a model is deployed in practice, the finally estimated function depends on the available sample. Therefore, the model should be selected, whose current function estimate on the given sample achieves the best score. Thus, the selection

[7] The descriptions as "priori" and "posteriori" perspective of evaluation should not be confused with the usual meanings in the context of Bayes statistics, but should be independently understood.

[8] The definition is equivalently transferable to the error probability.

of models is actually a selection over concrete functions. This means, that the model selection is ideally conditional on the concrete sample. Hence, an additional way of evaluation is considered based on an "a-posteriori", i.e. sample-wise or local selection. For a model class, for each of the 200 samples the best score across the tested models is focused. Different model classes are then compared based on the distribution of these conditionally best values. In this way no single models are selected, but the model class as a whole is represented by the series of minimal values over the set of samples. By calculating the average of these posteriori scores, the following value is estimated in the case of the L^2 distance to λ: $\mathbb{E}(\min_{\gamma_j,\delta_j} l_2(\hat{\lambda}_{\gamma_i,\delta_i}((Y_i, X_i)^{(l)})))$ [9]. Hence, a model class is represented by the expected minimum score over the corresponding models, as contrast to the priori evaluation, which considers the minimal expected score. Furthermore, it holds $\min_{\gamma_j,\delta_j} l_2(\hat{\lambda}_{\gamma_j,\delta_j}((y_i, x_i)^{(l)})) \leqslant l_2(\hat{\lambda}_{\gamma((y_i,x_i)^{(l)}),\delta((y_i,x_i)^{(l)})})$ for any empirical hyperparameter selection algorithm $\gamma((y_i, x_i)^{(l)})$, $\delta((y_i, x_i)^{(l)})$ (e.g. cross-validation). Therefore, this kind of selection represents the theoretically best selection process. In this sense, the posteriori evaluation might be interpreted as applying a best-case analysis. Note, that this selection method is only feasible, because the scores $l_2(.)$ as well as $err(.)$ could be calculated, due to the known distribution in this study. However, in practice this is not possible. In addition, this selection process might be compared with the aforementioned structural risk minimization (sec. 2.2.4). There the model selection is not based on estimates of the risk of a model, but on an approximation of the score of the current estimated function of the given sample. This approximation is specified in terms of an upper bound. Related to here, this means: $\min_{\gamma_j,\delta_j} B_j$ with $B_j > l_2(\hat{\lambda}_{\gamma_i,\delta_i})$ with a probability $\eta(l)$. This is similar to the posteriori evaluation, except that instead of an upper bound the actual score of the estimated function can be calculated, due to the known data distribution. The respective expectation is therefore conceptually similar to the risk of the SRM-estimator. However, it should be clear, that here no structure in the sense of the SRM was defined over the γ, δ parameters. This is why, it is not a structural risk minimization in the usual sense. On the other hand, the priori evaluation might be compared with applying the moving-control estimator (sec. 2.2.4).

[9] Analogue definition for the error probability.

Information presentation

For each type of evaluation-metrics, the model classes are evaluated based on these two perspectives. Presented are for each model class the (marginal) averaged scores of the selected priori models as well as the average of the posteriori minimum values. However, to assess the effect of the $\mathbb{V}$-matrix and the invariants, the sample-wise differences of the scores between two models are analysed. In the priori evaluation, this means the average difference between the selected models of two model classes. In the posteriori case, it is the average difference of two minimum score sequences of two model classes. Only model classes differing in one technicality are paired up, to evaluate the immediate effect of the different concepts. Because all models are estimated on the same set of samples, this is a joint sample evaluation. To facilitate the interpretation of the calculated scores, they are rescaled similar to the original paper. The L^2 distances are scaled by the L^2 norm of the conditional probability function. The error probabilities are transformed to relative quantities, i.e. $\tilde{err} = \frac{err}{err_{bayes}}$. Here err_{bayes} is the lowest achievable probability of false classification among any decision rules (sec. 2.2.1). Thus, the rescaled quantity is always greater than 1.

Assess uncertainty

To assess the uncertainty in the calculated mean estimates confidence intervals are provided. However, their calculation is accompanied by some complications. The provided confidence intervals are calculated based on the normality assumption, according to the central limit theorem. One reason, which supports this assumption is the comparatively high number of samples (200), such that a reasonable convergence might be plausible. To assess the potential convergence, the estimated skewness and kurtosis of the respective scores (-differences) are provided. What might question the normal assumption is, that the average values of the selected models in case of the priori analysis are minimas within the respective model class. Therefore, the averages should rather be treated by an extreme value distribution. Depending on how variable the selection of the same model is, this might vary from the normal distribution. To overcome this problem, these best models were evaluated again on an independent set of 200 samples. The presented quantities in the priori evaluation are based on these runs. To still provide an alternative to the average based comparisons, the evaluation is supplemented by quantile based information, specifically median estimates. This has the advantage, that for quantiles exact, nonparametric, non-trivial confidence intervals can be calculated [20], which are applied. Another problem is the effect of multiple testing. When analysing the pairwise average differences, several confidence intervals are calculated. To weaken the effect of false significance, the used cover probability is increased to 0.9995, such that the associated significance tests are equivalent to an error probability of 0.005. Also, the number of comparisons is hold as small as possible.

4.4 Results

In the following the obtained results are presented and conclusions made. The evaluation focuses on these five questions:

1. Is there an improvement by using the $\mathbb{V}$-matrix in the function approximation or the classification quality?
2. Is there an improvement by using statistical invariants in the function approximation or the classification quality?
3. Is there an improvement by using the modified version of reweighting the $\mathbb{V}$-matrix in the classification accuracy?
4. Is there any difference observable when using different sets of invariants?
5. Does the relaxation of the statistical invariants, by allowing deviations affect the performance?

The first three questions shall be answered regarding statistical significance.

4.4.1 Marginal performances

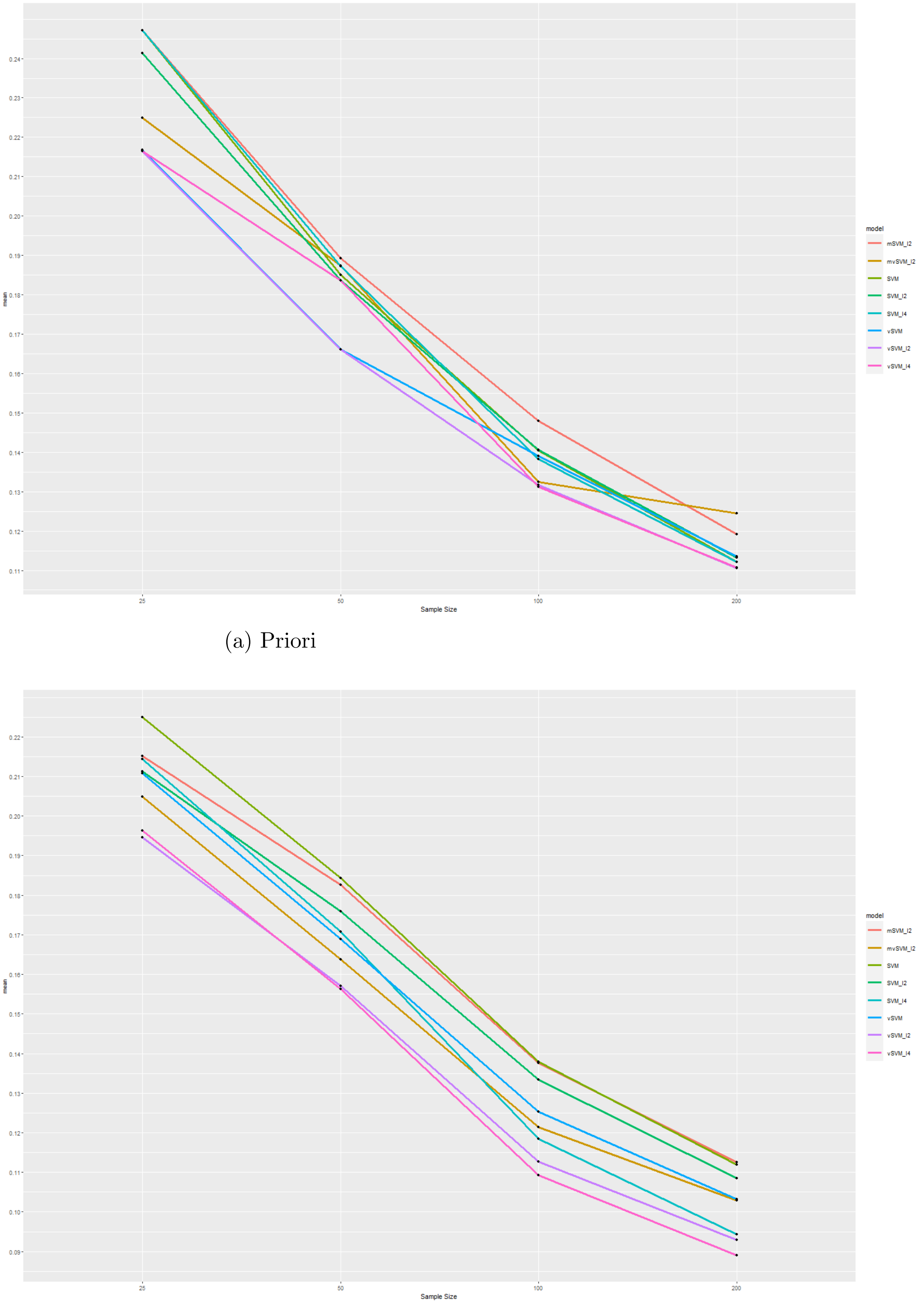
To get a first overview, the average values according to the priori and posteriori evaluation of the various model classes are displayed depending on the sample size. Because the very concrete values are of not much interest at this point, no confidence intervals are provided. This kind of information is comparable with the one given in the original paper.

Function approximation of λ

The lineplots of figure 4.1 show the relative L^2 distances. The smaller this value, the better is the approximation. The L^2 norm of λ is about 0.53253. Although for the evaluation of the function approximation the modified version of reweighting the $\mathbb{V}$-matrix is of not much relevance, it is still included for the sake of completeness, as was done in the original paper. At first, one confirms an intuitive effect, which is that for all models the performance improves with increasing sample size.

Priori evaluation

In figure 4.1(a) the constantly best model across all samples sizes is the complete LUSI method vSVM_I2. The models considering the $\mathbb{V}$-matrix (vSVM, vSVM_I2, vSVM_I4) represent the three lowest lines and thus the best performances. This might already indicate an advantage of the inverse-problem based approach for function estimation. Especially for the smaller sample sizes, the models realizing the regularized squared-loss minimization (SVM, SVM_I2, SVM_I4) lead to comparatively high L^2 distances. On the other hand, these show the steepest improvement over the sample size. Furthermore, the models SVM, SVM_I2, SVM_I4 perform very similar across all sample sizes. This grouping is analogue to the corresponding methods considering the $\mathbb{V}$-matrix. This might indicate, that the invariants and especially the concrete set of invariants are not of much relevance. Although the mvSVM_I2 approach is at first in the middle field, which might underline the importance of an initially better function estimate, due to the $\mathbb{V}$-matrix approach. The reweighting of the observations, represented by the models mSVMI2, mvSVM_I2, ends up causing the worst models for $l = 200$. While in the beginning the maximal spread between any of the models is about 0.03, the overall difference gets smaller with increasing sample size down to approximately 0.015. However, this magnitude leads to the conclusion, that despite some differences the approximations of the different models are actually of very similar quality.



(a) Priori

(b) Posteriori

Figure 4.1: The marginal performances of the various model classes with respect to the approximation of the conditional probability function dependent on the sample size. (a) Priori perspective: Averaged scaled L^2 distance of the best models per model class. Best models were evaluated on a second set of independent samples. (b) Posteriori perspective: Average of the smallest scaled L^2 distance per sample for each model class.

Posteriori evaluation

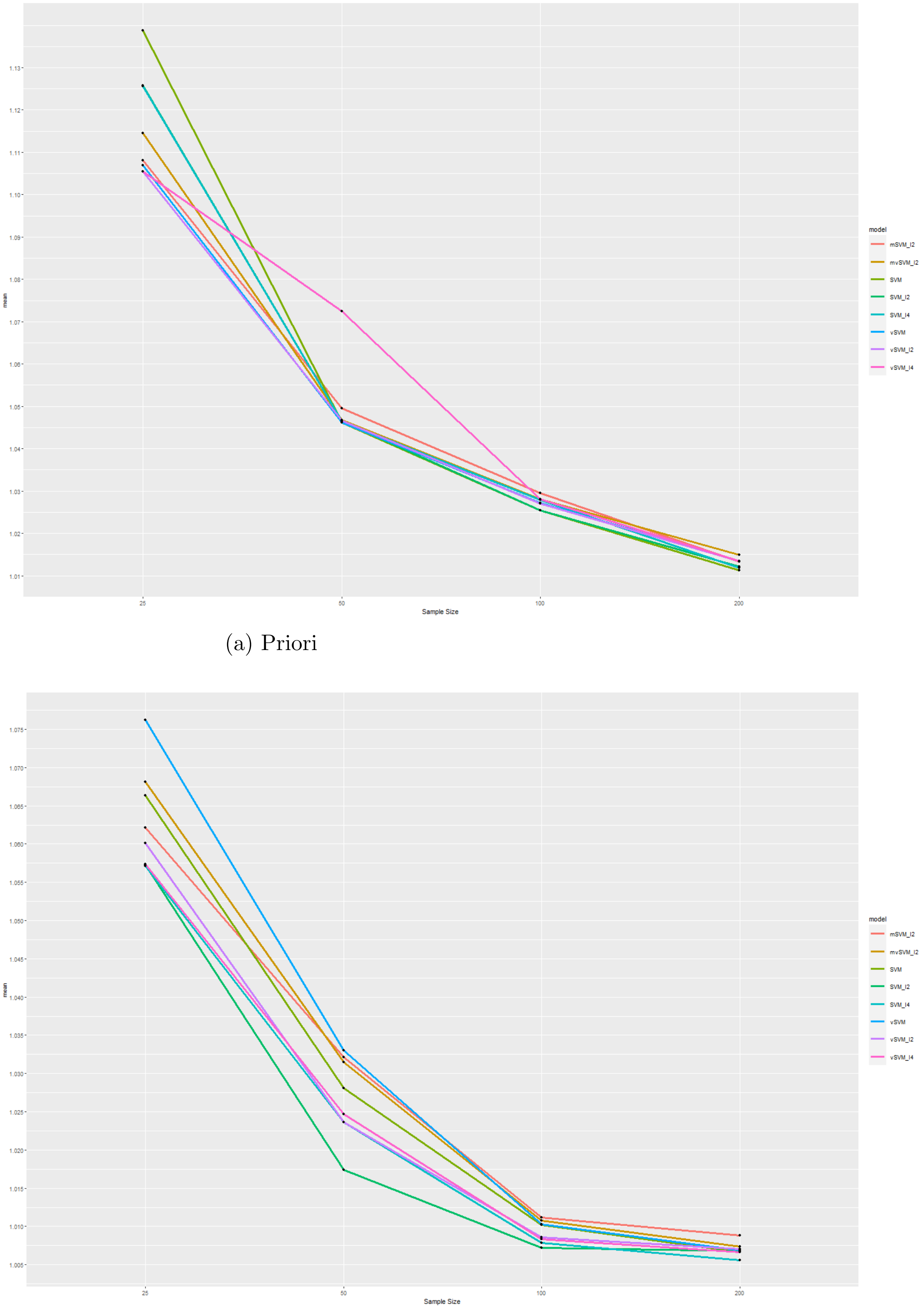
The individual models achieve a smaller L^2 distance than in the priori case (fig. 4.1(b)). This is obvious, because on each sample the smallest score value across a model class is selected. Again, the best models are the vSVM_I2 and vSVM_I4, at which vSVM_I4 ends up being slightly better for $l = 200$. The worst two are the regularized squared-loss estimates of SVM and the modified version mSVM_I2. Comparing models with respect to the application of the $\mathbb{V}$-matrix it seems, that the $\mathbb{V}$-matrix associated approach of approximating the conditional probability function performs better than the models following the empirical risk minimization of the regularized squared-loss (compare SVM : vSVM, SVM_I2/4 : vSVM_I2/4). Analysing the influence of the invariants it seems, that incorporating the statistical invariants leads to a better performance than their corresponding models ignoring them (compare SVM : SVM_I2/4 , vSVM : vSVM_I2/4). Furthermore, the posteriori evaluation might lead to the conclusion, that the larger set of invariants (..._I4) entails a slightly better approximation, if the sample size is big enough. Differently to the priori case, there is not much crossing between the lines. The lines stay parallel, thus the improvement over the sample size is for all models constant. This does also mean, that there is no kind of interaction effect between the application of the $\mathbb{V}$-matrix and the statistical invariants. The maximum difference does not decrease and stays about the same of roughly 0.03.

Classification performance

The figure 4.2 shows the respective relative error probability of the estimated functions' associated decision rules. The smaller the value, the better the performance. The Bayes error is about 0.15124. This value is actually similar to the documented value in the original paper, which carefully indicates that the data distribution was not too differently specified. As for the L^2 distance, all performances become better when increasing the sample size.

Priori evaluation

At first one recognizes, that for the sample sizes, which were considered in the original paper, i.e. $l \geqslant 50$, the classification quality is almost identical among the models. Therefore, it is difficult to carve out any (reliable) differences based on figure 4.2(a). For $l \geqslant 50$ the total difference among all performances is about 0.005, which shows that these differences can not be of any practical relevance. Taking a close look, it seems like the models considering the $\mathbb{V}$-matrix perform slightly worse for $l \geqslant 100$. While the models of SVM and SVM_I2/4 are similarly best at $l = 100$ and $l = 200$. This pattern also suggests, that the application of statistical invariants is not affecting the classification. Especially the modification of reweighting the $\mathbb{V}$-matrix does not establish any relevance. Interestingly, the models seem to exchange their performance ranking for the case of $l = 25$. It holds the following. The models realizing the $\mathbb{V}$-matrix based approach of function approximation do perform best. Surprisingly, the method SVM of regularized squared-loss minimization, which is more common in the context of classification, performs worst and is identical with its correspondent SVM_I2. The model of SVM_I4 achieves a smaller error probability than the analogue models with less invariants (SVM, SVM_I2). For the approaches considering the $\mathbb{V}$-matrix (vSVM, vSVM_I2/4) the incorporation of the invariants does not lead to any difference. Therefore, no effective influence of the invariants can be concluded also concerning the choice of invariants. The modifications by reweighting (mSVM_I2, mvSVM_I2) perform similar to vSVM and do not cause any improvement. The biggest difference at $l = 25$ between any of the models is about 0.035.



(a) Priori

(b) Posteriori

Figure 4.2: The marginal performances of the various model classes with respect to the classification dependent on the sample size. (a) Priori perspective: Averaged relative error probability of classification of the best models per model class. Best models were evaluated on a second set of independent samples. (b) Posteriori perspective: Averaged relative error probability of the minimum values per sample for each model class.

Posteriori evaluation
Similar to the L^2 distances, the lines of the posteriori evaluation do not cross as often as in the priori evaluation (fig. 4.2(b)). The maximum difference at $l = 25$ is even smaller than for the priori case of about 0.015, which differently to the priori case is also hold up for $l = 50$. Afterwards, the total differences shrink similar to the priori case below 0.005. It seems, that the application of the invariants is in fact decisive and improving the performance, independently of the use of the $\mathbb{V}$-matrix. The model SVM_I2 is constantly the best, especially for $l \leqslant 50$, followed by the models of vSVM_I2, vSVM_I4, SVM_I4, which perform almost identically. Surprisingly, the model vSVM is worst for small sample sizes, different to the priori evaluation. Hence, at this point the LUSI based approach of function approximation can not be claimed to improve the classification. The modified methods of reweighting (mSVM_I2, mvSVM_I2) perform similarly to the straight empirical risk minimization model SVM. As for the L^2 distance, the posteriori evaluation indicates a weak tendency of improvement, when incorporating the statistical invariants. But again, taking the concrete magnitude into account it is obvious, that all models perform already quite well. Especially for the larger sample sizes the Bayes error is almost achieved, as the values are near to 1.

Evaluation of δ values

As explained, different values of admissible deviation in the linear side constraints of the statistical invariants were tested within the respective model classes (tab. 4.1). The value range also included zero, meaning that strict equality is demanded. To detect whether allowing these deviations has made any difference, the delta values of the best models concerning the priori evaluation for the various model classes considering invariants are shown in table B.2 in the appendix. The table clearly shows, that for both, the quality of function approximation, i.e. L^2 distance to λ, as well as for the classification accuracy these selected models are mostly allowing some deviation in the statistical invariants. Although it might be, that the effect difference to the best models of strict equality is negligible, any deeper analysis is left out at this point, as in practice incorporating these hyperparameters is rather convenient.

4.4.2 Pairwise evaluation of effects

To analyse the effect differences of the various models in greater detail, the following pairwise comparisons are given. In the subsequent plots, the first three pairs should be consulted to assess the effect of the $\mathbb{V}$-matrix, while the last four are dealing with the influence of deploying the invariants. The two middle ones, fourth and fifth, are about evaluating the modified reweighting approach. Note, that a positive value means that the first method with respect to the labels performs worse. The plots contain the comparisons for the four different sample sizes, denoted in the labels by "$\#l$".

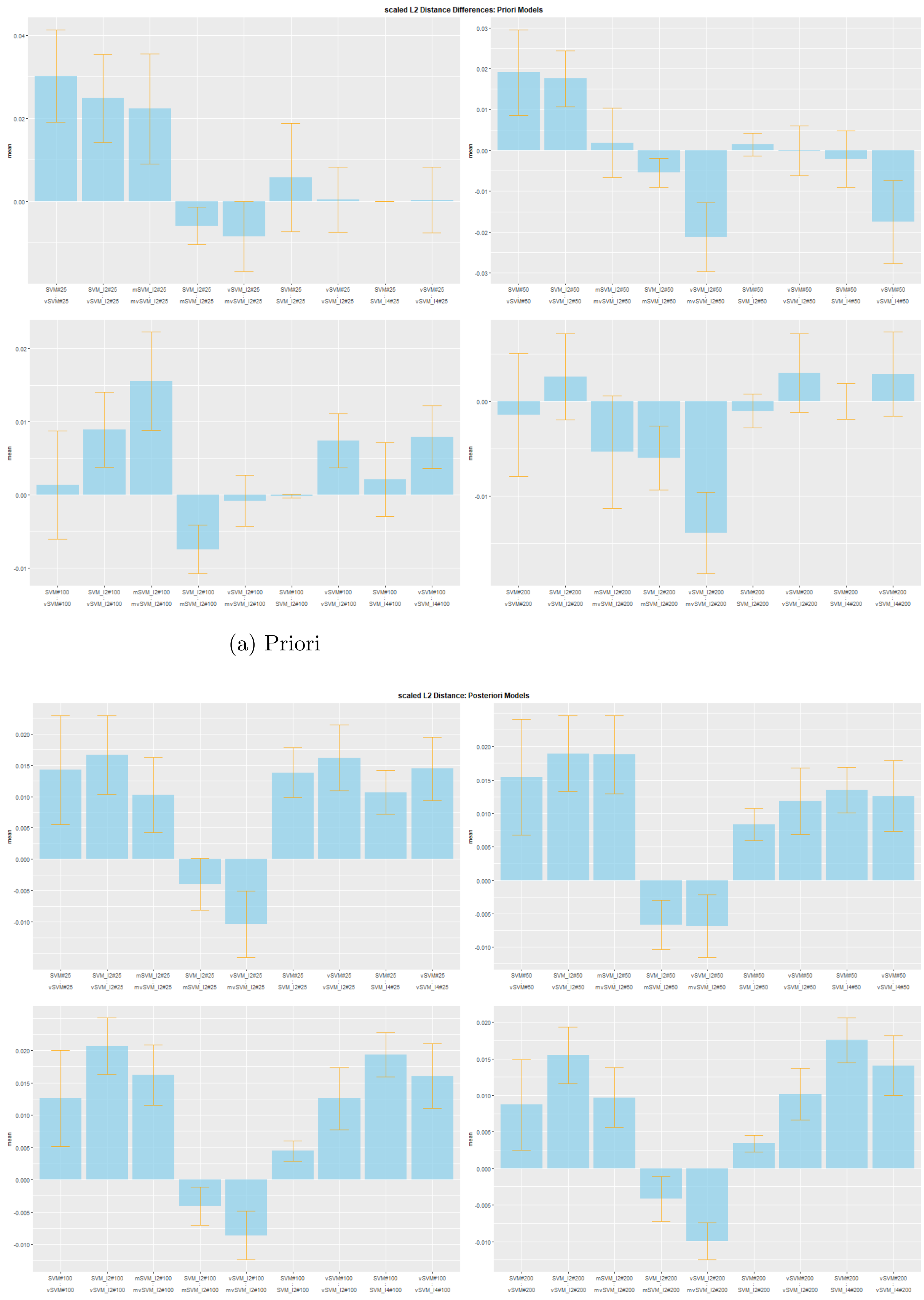
Function approximation of λ

Priori evaluation

$\mathbb{V}$-matrix

Pairs 1, 2 (SVM : vSVM, SVM_I2 : vSVM_I2) fig. 4.3(a): For small sample sizes the $\mathbb{V}$-matrix leads indeed to a significant improvement. Also, for the case of $l = 100$ the effect of the complete LUSI approach vSVM_I2 is still significant against SVM_I2. On the other hand, it becomes clear that as the sample size increases this impact vanishes, beginning with an effect difference of about 0.03 down to less than 0.0025. For $l = 200$ the effect differences are not significant anymore. This observation seems plausible as with more data, the empirical regularized squared-loss minimizer should estimate the conditional probability function precisely as well.

Pair 3, (mSVM_I2 : mvSVM_I2): For models considering the reweighting, the influence of the $\mathbb{V}$-matrix is not really consistent. While for $l = 25$ and $l = 100$ it shows a comparatively high significant improvement, the effect difference almost vanishes for $l = 50$ and $l = 200$ and even worsens the quality.



(a) Priori

(b) Posteriori

Figure 4.3: The pairwise comparisons of model classes regarding the approximation of the conditional probability function. (a) Priori perspective: Averaged differences of scaled L^2 distances between the best models of two model classes. Best models were evaluated on a second set of independent samples. (b) Posteriori perspective: Averaged differences of the minimum values of scaled L^2 distances per sample of two model classes.

Invariants

Pairs 6 - 9 fig. 4.3(a): The invariants show instead along all sample sizes and methods no real improvement. The effect differences do almost vanish and are in most cases not even significant. For $l = 100$ a small improvement might be observable. However, in general the impact on the function estimation seems rather negligible. Accordingly, the concrete set of invariants does not make any difference.

Reweighted $\mathbb{V}$-matrix

Pairs 4, 5 (SVM_I2 : mSVM_I2, vSVM_I2 : mvSVM_I2) fig. 4.3(a): The proposed modification based on adjusting the $\mathbb{V}$-matrix does impair the function approximation. The evaluation indicates that this negative effect is stable along the various sample sizes and does not vanish. It becomes significant for $l = 50, 200$. This is intuitive, because the modification was proposed to benefit the classification and not the function approximation in the first place.

Posteriori evaluation

$\mathbb{V}$-matrix

Pairs 1, 2 (SVM : vSVM, SVM_I2 : vSVM_I2) fig. 4.3(b): The conclusions are similar to the priori evaluation. The $\mathbb{V}$-matrix based approach of function approximation establishes a significant improvement. Other than in the priori evaluation, the effect size stays about the same along the sample sizes as already observed in the previous line-plots (fig. 4.1). The effect size is comparable to the priori evaluation and is in all cases less than 0.02.

Pair 3, (mSVM_I2 : mvSVM_I2): Other than in the priori case, also for the reweighting based models the $\mathbb{V}$-matrix holds up a significant improvement.

Invariants

Pairs 6 - 9 fig. 4.3(b): Other than in the priori evaluation, the invariants based methods imply in all cases a significant improvement. For the models considering only two invariants (compare SVM : SVM_I2, vSVM : vSVM_I2), the influence seems to be bigger when the model applies the $\mathbb{V}$-matrix. The effect differences begin at 0.015 and decrease with increasing sample size down to 0.005 or 0.01 respectively. Thus, end being rather negligible. In case of four invariants, their influence is greater for the regularized squared-loss minimization based models (compare SVM : SVM_I4, vSVM : vSVM_I4). The effect size increases with increasing sample size up to 0.0175.

Reweighted $\mathbb{V}$-matrix

Pairs 4, 5 (SVM_I2 : mSVM_I2, vSVM_I2 : mvSVM_I2) fig. 4.3(b): Likewise the priori evaluation, the modification tends to worsen the function approximation. Although the effect difference is only small, it is significant and negative for both pairs and all sample sizes.

Median informed evaluation

In table B.1 in the appendix, the estimated skewness and kurtosis of the difference value distributions are given, to help assess the reasonableness of the normal assumption of the respective average values. While the skewness is in many cases reasonably close to zero, the kurtosis is especially for the models considering the invariants rather big, indicating the presence of some extreme observations. Therefore, to supplement the evaluation, the same comparisons based on the estimation of the median differences are provided [10]. The corresponding plots can be found in the appendix figure B.3. The findings can be summarized as follows.

Priori evaluation

Considering this quantile information, it is recognized, that the results are quite similar, which gives the general evaluation more credibility (fig. B.3(a)). The statistical significance and the magnitudes are in most cases identical. Slightly different is the somewhat more pronounced statement with respect to the invariants. Here the plot reveals the presence of some extreme observations as the median is mostly smaller and sometimes set to almost 0. Thus, it even further lessens the influence of the invariants.

Posteriori evaluation

The statements about the statistical significance are identical. All observed differences basically remain significant. In most cases the effect difference is even smaller. The effect of the $\mathbb{V}$-matrix seems to become bigger with increasing sample size and converges for the respective comparisons towards 0.01. Other than for the average based evaluation the effect of the invariants shrinks with increasing

[10] Note, that the determination of a best model for a model class in case of the priori evaluation was still done on the basis of the average values. The calculated medians for the priori evaluation are these of the evaluation of the best models on the new set of samples.

sample size. This is especially for the models involving the set of two invariants, whose differences are almost zeroed. One exception is SVM : SVM_I4, which remains similar to the previous analysis and grows to 0.015. The modification of reweighting is not showing any improvement. In fact, with increasing sample size a small decline is observable.

Classification performance

Priori evaluation

$\mathbb{V}$-matrix

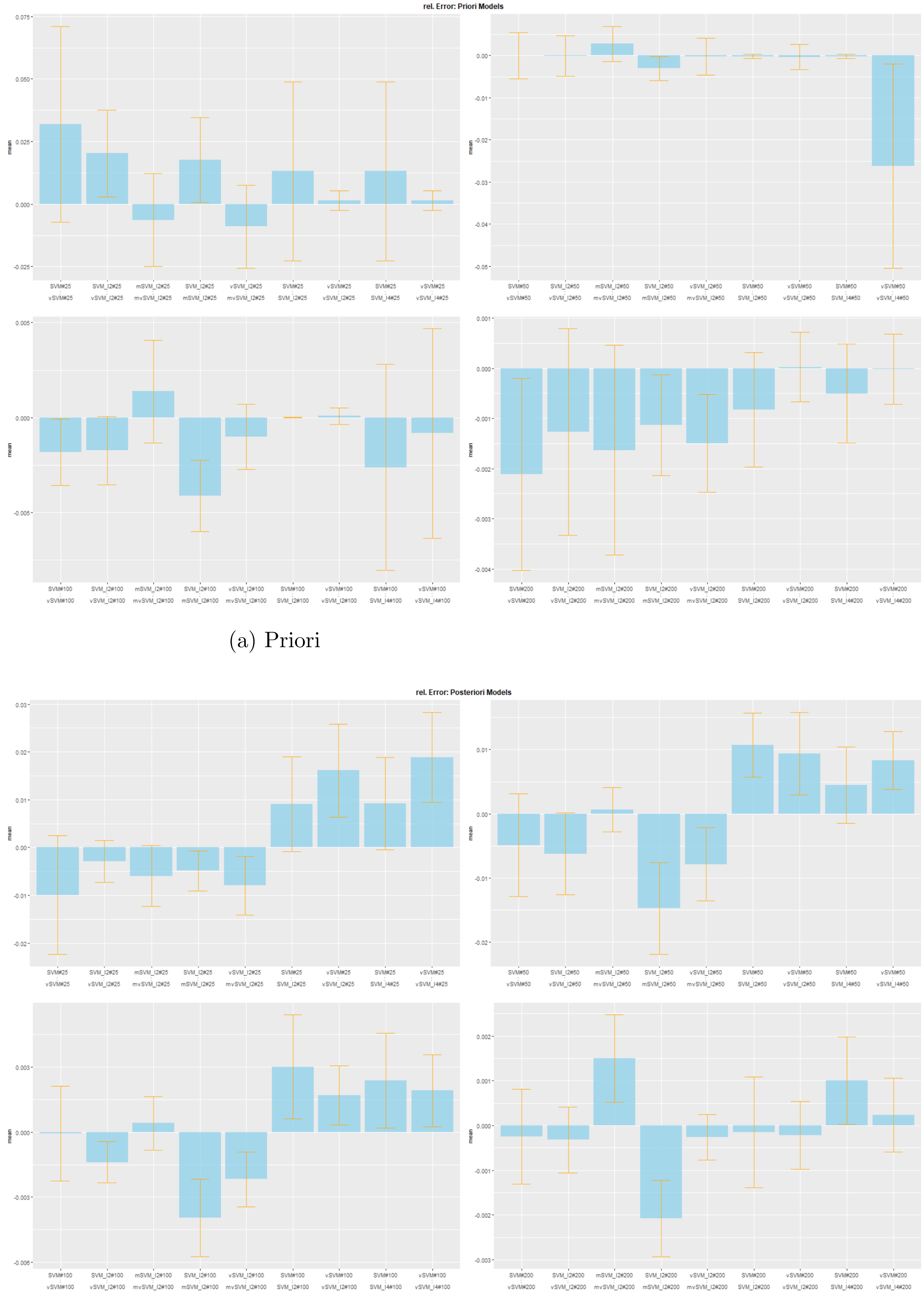
Pairs 1, 2 (SVM : vSVM, SVM_I2 : vSVM_I2) fig. 4.4(a): The comparisons show, that the LUSI based approach of estimating the conditional probability function never establishes a significant improvement. Increasing the sample size to $l = 100, 200$, the $\mathbb{V}$-matrix actually seems to worsen the classification ability, sometimes even significantly. On the other hand, the magnitude of these differences is at its most about -0.002 ($l = 200$) and therefore out of any practical relevance.
Pair 3, (mSVM_I2 : mvSVM_I2): Also, the reweighting does not benefit from deploying at first the $\mathbb{V}$-matrix. Instead, the effects are always insignificant.

Invariants

Pairs 6 - 9 fig. 4.4(a): Incorporating the invariants does not show any significant difference compared to ignoring them. The effect differences are negligible and with increasing sample size it rather worsens the accuracy.

Reweighted $\mathbb{V}$-matrix

Pairs 4, 5 (SVM_I2 : mSVM_I2, vSVM_I2 : mvSVM_I2) fig. 4.4(a): Although the idea of reweighting the objective function targets the classification, no significant improvement can be established. Instead, by increasing the sample size a significant decline can be observed.



(a) Priori

(b) Posteriori

Figure 4.4: The pairwise comparisons of model classes regarding classification. (a) Priori perspective: Averaged differences of the relative error probability between the best models of two model classes. Best models were evaluated on a second set of independent samples. (b) Posteriori perspective: Averaged differences of the minimum relative error probabilities per sample of two model classes.

Posteriori evaluation
In general the conclusions are about the same as for the priori evaluation (fig. 4.4(b)).

$\mathbb{V}$-matrix
Pairs 1, 2 (SVM : vSVM, SVM_I2 : vSVM_I2) fig. 4.4(b): Again, the effect differences are marginal and not significant. But for all sample sizes, the average value is negative, indicating a small increase in the error probability when following the function approximation approach according to the $\mathbb{V}$-matrix.
Pair 3, (mSVM_I2 : mvSVM_I2): In case of the modified alternative of reweighting, the positive influence of the $\mathbb{V}$-matrix increases with increasing sample size. For $l = 200$ even a significant improvement can be observed. However, the effect difference is only about 0.0015.

Invariants
Pairs 6 - 9 fig. 4.4(b): Differently than for the priori evaluation, the invariants based methods do show some improvement as well as significant influence for the sample sizes $l = 25, 50, 100$. In general, the effect size does shrink with increasing sample size. The most pronounced effect is for vSVM_I4 in case of $l = 25$ of roughly 0.02.

Reweighted $\mathbb{V}$-matrix
Pairs 4, 5 (SVM_I2 : mSVM_I2, vSVM_I2 : mvSVM_I2) fig. 4.4(b): Again, the reweighting does lead to an increase in the error probability, especially for the models based on the regularized squared-loss minimization. For all sample sizes the average effect is even significant.

The estimated differences in the error probabilities are of small magnitude, which is why even a significant difference is out of any practicality here. This must be kept in mind, when interpreting the observed differences. Thus, the practical relevance for the classification is mostly clarified. This is why, no evaluation based on quantile information as for the function approximation is provided here.

4.5 Conclusion

4.5.1 Summarizing and relating findings

The conclusions of the original paper [11] have to be treated carefully as they are obtained on the basis of a single sample and no uncertainty quantification is provided. The authors argue that both parts, i.e. the $\mathbb{V}$-matrix as result of the function approximation in terms of the inverse-problem definition of the conditional probability function and the Tikhonov regularization, as well as the statistical invariants lead to explicit improvements. These statements could not be confirmed with the same generality and clearness.

Function approximation of λ

$\mathbb{V}$-matrix

Across all evaluations most often a clear, significant improvement could be observed. While for the posteriori evaluation the effect difference remains similar or is slightly increasing with increasing sample size, the effect seems to vanish with increasing sample size considering the priori evaluation. This does actually make sense, as this model was intentionally constructed with focus on the function estimation. But with increasing sample size, the empirical risk minimizer of the regularized squared-loss might compensate this difference, whose limit is the conditional probability function too. The conclusions basically accord with the original paper, although the discovered effect in this thesis is much less expressive. The absolute magnitude is even at its most still below 0.03 (priori, $l = 25$).

Invariants

Deriving conclusions about the impact of the applied statistical invariants is somewhat more difficult, as it depends on the evaluation perspective. For the priori evaluation the invariants do not establish any clear improvement. Only single cases, especially the one SVM_I4 are constantly showing a significant (small) improvement. However, the most comparisons show a negligible and mostly insignificant difference. Furthermore, no real difference between the sets of invariants could be identified. In the posteriori evaluation a significant, small yet constant improvement is always present, which is at its most still below 0.02 (posteriori,

$l = 100$). In the posteriori evaluation a fine difference might be identified, that the set of four invariants has a more stable positive effect than the set of two invariants. The median based evaluation does underline these results, although the effect size is usually somewhat smaller such that the invariants are of no practical use then anymore. Summarizing, a distinct and systematic improvement by the statistical invariants could not be proven for the considered data distribution. This is differently depicted in the original study. However, it should be noted, that clearly the construction of predicates and corresponding invariants was not exhaustively tested here. Furthermore, the posteriori evaluation gives at least some indication, that the concept of statistical invariants should be investigated in the concrete use case. Especially because the influence of the invariants can conveniently be controlled by the allowed deviation δ, which is why the models of strict equality, as well as discarding the invariants are implicitly contained.

Reweighted $\mathbb{V}$-matrix

The modification of adjusting the $\mathbb{V}$-matrix does not benefit the function estimation at all. Instead, its influence seems indifferent and in some cases even worsens the performance, especially with increasing sample size. Considering that reweighting the $\mathbb{V}$-matrix served the purpose of classification this result is not very surprising. However, the conclusion drawn here is opposite to the statement in the original paper.

Classification performance

The investigations show, that basically none of the introduced LUSI related methods have any relevant improving impact on the classification performance. This is directly proven by the small magnitude of effect differences, especially when being aware of the fact, that these are relative values (sec. 4.3). All evaluated models perform almost perfectly. Only in the priori evaluation the methods vSVM and vSVM_I2 seem to have a small improving influence for $l = 25$. Otherwise, they rather tend to worsen the classification accuracy in both the priori and posteriori evaluation. Furthermore, the modification of reweighting the $\mathbb{V}$-matrix does not imply any improvement. Consulting the posteriori evaluation, a small benefit is shown for small sample sizes when involving the invariants. These conclusions contradict the claimed performance boosts for classification in the original paper. The evaluation of L^2 distance and classification accuracy at the same time underlines the role of defining the concrete risk functional and how performances are differently reflected. Although knowing the conditional probability function would mean having a much richer set of information, the immediate goal is to minimize the probability of false classification for the pattern recognition problem. Therefore, a more direct approach such as the empirical risk minimization of the squared-loss is due to the estimation variation at least equally successful. Hence, the obtained results conform with the theory.

4.5.2 Outlook

The results show, that all models achieve almost the perfect Bayes error, especially for $l \geqslant 50$. This indicates, that the chosen example might be too simple to actually carve out the differences between the methods regarding the pattern recognition problem. Therefore, a new evaluation on a more complex distribution might be meaningful. Such a distribution could already be defined by a conditional probability function with multiple cross points at the level of 0.5. Also, a uniform distribution over $\mathcal{X}$ might be reasonable, to get more observations at the extremes. This would avoid, that the effect differences only arise because of the methods' (insufficient) performance at the scope's limits. Furthermore, one might elaborate the analysis of the statistical invariants. Because this study showed in single cases some improving trends, an in-depth analysis might be promising. Such an investigation could focus on both. On the one hand, the theoretical analysis of defining meaningful predicate functions, also for a multi dimensional domain $\mathcal{X}$. In the original paper some first approaches are already given. On the other hand, an empirical evaluation could be conducted in greater detail. This might entail the construction and evaluation of predicate selection algorithms and how different subsets of predicates influence the performance. Finally, to assess the models under realistic circumstances, an evaluation based on real datasets should be considered. Similar tests were already given in the original paper. However, again most of the information of the experimental setup was not documented. Furthermore, only the methods SVM and vSVM_I2 were evaluated, which is why clear statements about the individual effects of neither the $\mathbb{V}$-matrix nor the invariants are deducible.

Bibliography

[1] S. Kabanikhin, “Definitions and examples of inverse and ill-posed problems,” *Journal of Inverse and Ill-posed Problems*, vol. 16, no. 4, pp. 317–357, 2008, doi:10.1515/JIIP.2008.019.

[2] Springer and European Mathematical Society, “Well-posed problem,” https://encyclopediaofmath.org/wiki/Well-posed_problem, [Online; accessed 16-May-2022].

[3] V. Vapnik, *Estimation of dependences based on empirical data.* Springer, 1982, translated by S. Kotz.

[4] ——, *Estimation of dependences based on empirical data; Empirical inference science, Afterword of 2006.* Springer, 2006.

[5] M. Kokurin, “Conditionally well-posed and generalized well-posed problems,” *Computational Mathematics and Mathematical Physics*, vol. 53, no. 6, pp. 681–690, 2013, doi:10.1134/S0965542513060110.

[6] A. Tikhonov, “Tikhonov’s work on methods of solving ill-posed problems,” *Russian Mathematical Surveys*, vol. 22, no. 2, pp. 142–149, 1967, doi:10.1070/rm1967v022n02abeh001216.

[7] C. Clason and A. Klassen, “Quasi-solution of linear inverse problems in non-reflexive Banach spaces,” *Journal of Inverse and Ill-posed Problems*, vol. 26, no. 5, pp. 689–702, 2018, doi:10.1515/jiip-2018-0026.

[8] C. Groetsch, “Integral equations of the first kind, inverse problems and regularization: a crash course,” *Journal of Physics: Conference Series*, vol. 73, no. 1, 2007, doi:10.1088/1742-6596/73/1/012001.

[9] H. Chen and Z. Hou, "Modified discrepancy principles with perturbed operators and noisy data," *Journal of Computational Mathematics*, vol. 14, no. 2, pp. 108–119, 1996.

[10] F. Lenzen and O. Scherzer, "Tikhonov type regularization methods: History and recent progress," *Proceeding Eccomas*, vol. 2004, 2004.

[11] V. Vapnik and R. Izmailov, "Rethinking statistical learning theory: Learning using statistical invariants," *Machine Learning*, vol. 108, no. 3, pp. 381–423, 2018, doi:10.1007/s10994-018-5742-0.

[12] V. Vapnik and A. Y. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," *Theory of Probability and its Applications*, vol. 16, no. 2, pp. 264–280, 1971, doi:10.1137/1116025.

[13] U. von Luxburg and B. Schölkopf, "Statistical learning theory: Models, concepts and results," in *Inductive Logic*, ser. Handbook of the History of Logic, D. Gabbay, S. Hartmann, and J. Woods, Eds. Elsevier North-Holland, 2011, vol. 10, pp. 651–706, doi:10.1016/B978-0-444-52936-7.50016-1.

[14] F. Cucker and S. Smale, "On the mathematical foundations of learning," *Bulletin of the American Mathematical Society*, vol. 39, no. 1, pp. 1–49, 2001, doi:10.1090/S0273-0979-01-00923-5.

[15] V. Paulsen and M. Raghupathi, *An introduction to the theory of reproducing kernel Hilbert spaces*, ser. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2016, doi:10.1017/CBO9781316219232.

[16] D. Werner, *Funktionalanalysis.* Springer Spektrum, 2018.

[17] V. Vapnik and R. Izmailov, "Synergy of monotonic rules," *Journal of Machine Learning Research*, vol. 17, no. 136, pp. 1–33, 2016.

[18] GUROBI, "R-API overview," https://www.gurobi.com/documentation/9.5/refman/r_api_overview.html, [Online; accessed 16-May-2022].

[19] ——, "Guidelines for numerical issues," https://www.gurobi.com/documentation/9.5/refman/guidelines_for_numerical_i.html, [Online; accessed 16-May-2022].

[20] C. Ialongo, "Confidence interval for quantiles and percentiles," *Biochemia Medica*, vol. 29, no. 1, 2019, doi:10.11613/BM.2019.010101.

A Functional analysis

In the following some basic definitions of theory of functional analysis are provided. All definitions are obtained by [16].

Definition A.1 (Completeness, Banach space)**.** A metric space, at which each Cauchy sequence converges is called *complete.* A complete, normed vector space is a *Banach space.*

Definition A.2 (Completion of a metric space)**.** Let $(\mathcal{X}, d_{\mathcal{X}})$ be a metric space and $CS(\mathcal{X})$ the set of all Cauchy sequences of it. Then an equivalence relation over $CS(\mathcal{X})$ is defined: $(x_n) \sim (y_n) \iff d_{\mathcal{X}}(x_n, y_n) \to 0$. Let $\hat{\mathcal{X}}$ be the set of all equivalence classes and define: $d_{\hat{\mathcal{X}}}([(x_n)], [(y_n)]) := \lim_{n\to\infty} d(x_n, y_n)$. Then $(\hat{\mathcal{X}}, d_{\hat{\mathcal{X}}})$ is the completion of $\mathcal{X}$ and is a complete metric space, in which $\mathcal{X}$ lies densely.

This procedure can be applied canonically to a normed vector space. Its completion is then a Banach space.

Definition A.3 (Operator, Functional)**.** A function between two normed vector spaces is called an *operator.* If the image space is the scalar space, then it is called *functional.*

Definition A.4 (Operator norm)**.** For an operator $T : \mathcal{X} \to \mathcal{Y}$, define $\|T\| := \inf\{\, M \geqslant 0 : \|Tx\| \leqslant M\,\|x\| \;\forall\, x \in \mathcal{X}\,\}$. This defines on $L(\mathcal{X}, \mathcal{Y}) := \{\, T : \mathcal{X} \to \mathcal{Y} \mid \text{T is a linear and continuous operator}\}$ a norm and is called *operator norm.*

Proposition A.1. Let $\mathcal{X}$ and $\mathcal{Y}$ be normed vector spaces and $T : \mathcal{X} \to \mathcal{Y}$ be any linear operator. The following statements are equivalent: (1) T is continuous. (2) There exists a $M \geqslant 0$ s.t. $\|Tx\| \leqslant M\,\|x\|$. If T is continuous, it is also called *bounded.*

Definition A.5 (Isomorphism, Isometric)**.** A linear, continuous operator $T\colon \mathcal{X} \to \mathcal{Y}$ is called an *isomorphism*, if T is bijective and T^{-1} is continuous. If $\|Tx\| = \|x\|$ for all $x \in \mathcal{X}$, then T is called *isometric*.

Definition A.6 (Dual space)**.** The space $L(\mathcal{X}, \mathbb{K})$ is the set of continuous, linear functionals of a normed vector space $\mathcal{X}$ and is called the *dual space* $\mathcal{X}'$ of $\mathcal{X}$.

Corollary A.1. The dual space of a normed vector space together with the operator norm is always a Banach space.

Definition A.7 (Compact operator)**.** A linear operator T between $\mathcal{X}$ and $\mathcal{Y}$ is called *compact*, if $T(B_{\mathcal{X}})$ is relatively compact, i.e. $\overline{T(B_{\mathcal{X}})}$ is compact, at which $B_{\mathcal{X}} := \{x \,|\, \|x\| \leqslant 1\}$. The following statements are equivalent: (1) T is compact. (2) T maps bounded sets on relatively compact sets. (3) For each bounded sequence (x_n) in $\mathcal{X}$, the sequence $(Tx_n) \subseteq \mathcal{Y}$ has a convergent subsequence.
A compact operator is continuous.

Definition A.8 (Adjoint operator)**.** Let $\mathcal{X}, \mathcal{Y}$ be two normed vector spaces and $T \in L(\mathcal{X}, \mathcal{Y})$. Then the *adjoint* operator $T' : \mathcal{Y}' \to \mathcal{X}'$, is defined by $(T'y')(x) = y'(Tx)$. The adjoint operator is linear and continuous itself.

Definition A.9 (Inner product)**.** Let $\mathcal{X}$ be a $\mathbb{K}$ vector space. A function $\langle ., . \rangle : \mathcal{X} \times \mathcal{X} \to \mathbb{K}$ is called inner product (scalar product), if $\forall x, y, x_1, x_2 \in \mathcal{X}, \lambda \in \mathbb{K}$:

i $\langle x_1, x_2 \rangle = \langle x_1, y \rangle + \langle x_2, y \rangle$

ii $\langle \lambda\, x, y \rangle = \lambda \langle x, y \rangle$

iii $\langle x, y \rangle = \overline{\langle y, x \rangle}$

iv $\langle x, x \rangle \geqslant 0$

v $\langle x, x \rangle = 0 \iff x = 0$

Corollary A.2. The map $x \to \langle x, x \rangle^{1/2}$ defines a norm on $\mathcal{X}$.

Definition A.10 (Pre-Hilbert space, Hilbert space)**.** A normed vector space $\mathcal{X}$ is called *pre-Hilbert space*, if there is an inner product $\langle ., . \rangle$, such that $\|x\| = \langle x, x \rangle^{1/2}$. If $\mathcal{X}$ is complete, then it is a *Hilbert space*.

Definition A.11 (Orthogonality, Orthogonal complement)**.** Let $\mathcal{X}$ be a pre-Hilbert space. Two vectors $x, y \in \mathcal{X}$ are called *orthogonal*, $x \perp y$, if $\langle x, y \rangle = 0$. Two sets $A, B \subseteq \mathcal{X}$ are called orthogonal, $A \perp B$, if $x \perp y$ for all $x \in A$, $y \in B$. The set $A^{\perp} := \{y \in \mathcal{X} \,|\, x \perp y \, \forall \, x \in A\}$ is called *orthogonal complement* of A.

Proposition A.2 (Projection sentence)**.** Let $\mathcal{H}$ be a Hilbert space and $K \subseteq \mathcal{H}$ a closed, convex subset and $x_0 \in \mathcal{H}$. Then, there is exactly one $x \in K$ with: $\|x - x_0\| = \inf\limits_{y \in K} \|y - x_0\|$.

Theorem 1 (Representation theorem by Frechet-Riesz)**.** Let $\mathcal{H}$ be a Hilbert space. Then, the map $\Phi : \mathcal{H} \to \mathcal{H}', y \to \langle ., y \rangle$ is bijective, isometric and conjugate linear ($\Phi(\lambda \, y) = \overline{\lambda} \, \Phi(y)$). This means, that for any $x' \in \mathcal{H}'$ there is exactly one $y \in \mathcal{H}$, such that $x'(x) = \langle x, y \rangle$ for $x \in \mathcal{H}$ and $\|x'\| = \|y\|$.

Definition A.12 (Orthonormal system, Orthonormal basis)**.** A subset $S \subseteq \mathcal{H}$ is called an *orthonormal system*, if $\|e\| = 1$ and $\langle e, f \rangle = 0$ for $e, f \in S$, $e \neq f$. A orthonormal system S is called orthonormal basis, if: $S \subseteq P$, P orthonormal system $\implies P = S$.

Corollary A.3. Let $S \subseteq \mathcal{H}$ be a orthogonal system and $x \in \mathcal{H}$. Then, $S_x := \{e \in S \,|\, \langle x, e \rangle \neq 0\}$ is at most countable.

Definition A.13 (Unconditional convergence)**.** Let $\mathcal{X}$ be a normed vector space and I an infinite index set. Let $x_i \in \mathcal{X}$ for $i \in I$. The series $\sum_{i \in I} x_i$ converges *unconditionally* against $x \in \mathcal{X}$, if $I_0 = \{i \,|\, x_i \neq 0\}$ is at most countable and for any numeration $I_0 = \{i_1, i_2, ...\}$ the equation $\sum_{n=1}^{\infty} x_{i_n} = x$ holds. For infinite dimensional vector spaces, unconditionally convergence is not equivalent to absolute convergence anymore.

Proposition A.3. The following statements are equivalent: (a) S is an orthonormal basis. (b) $\forall x \in \mathcal{H} \, : \, x = \sum_{e \in S} \langle x, e \rangle \, e$, the convergence is unconditional.

Proposition A.4. For a Hilbert space $\mathcal{H}$ the following are equivalent: (a) $\mathcal{H}$ is separable (b) All orthonormal bases are countable. (c) There is one countable orthonormal basis.

Definition A.14 (Adjoint operator in Hilbert spaces)**.** Let $T : \mathcal{H}_1 \to \mathcal{H}_2$ be a linear, continuous operator between two Hilbert spaces. Let $\Phi_i : \mathcal{H}_1 \to \mathcal{H}_2'$ be the corresponding isomorphism of the Frechet-Riesz representation. The *adjoint* operator in the sense of Hilbert spaces is defined by $T^* = \Phi_1^{-1} T' \Phi_2$, at which T' is the original adjoint operator. Thus, it holds: $\langle Tx, y\rangle_{\mathcal{H}_2} = \langle x, T^* y\rangle_{\mathcal{H}_1}$.

Definition A.15 (Unitary, Self-adjoint, Normal)**.** T is called *unitary* if $T^{-1} = T^*$. Let $\mathcal{H}_1 = \mathcal{H}_2$. T is called *self-adjoint* if $T = T^*$. T is called *normal* if $T\,T^* = T^*\,T$.

B Supplementary

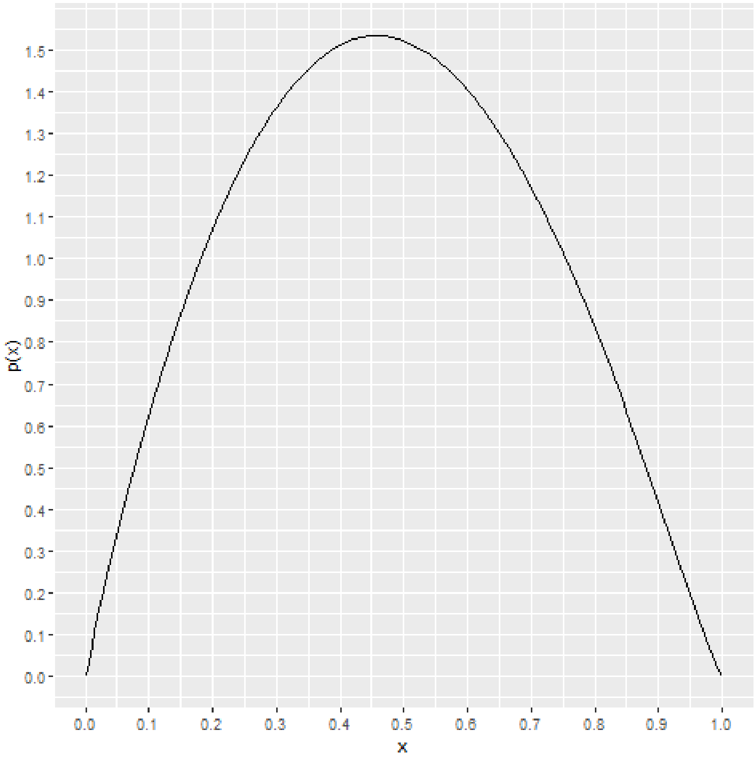
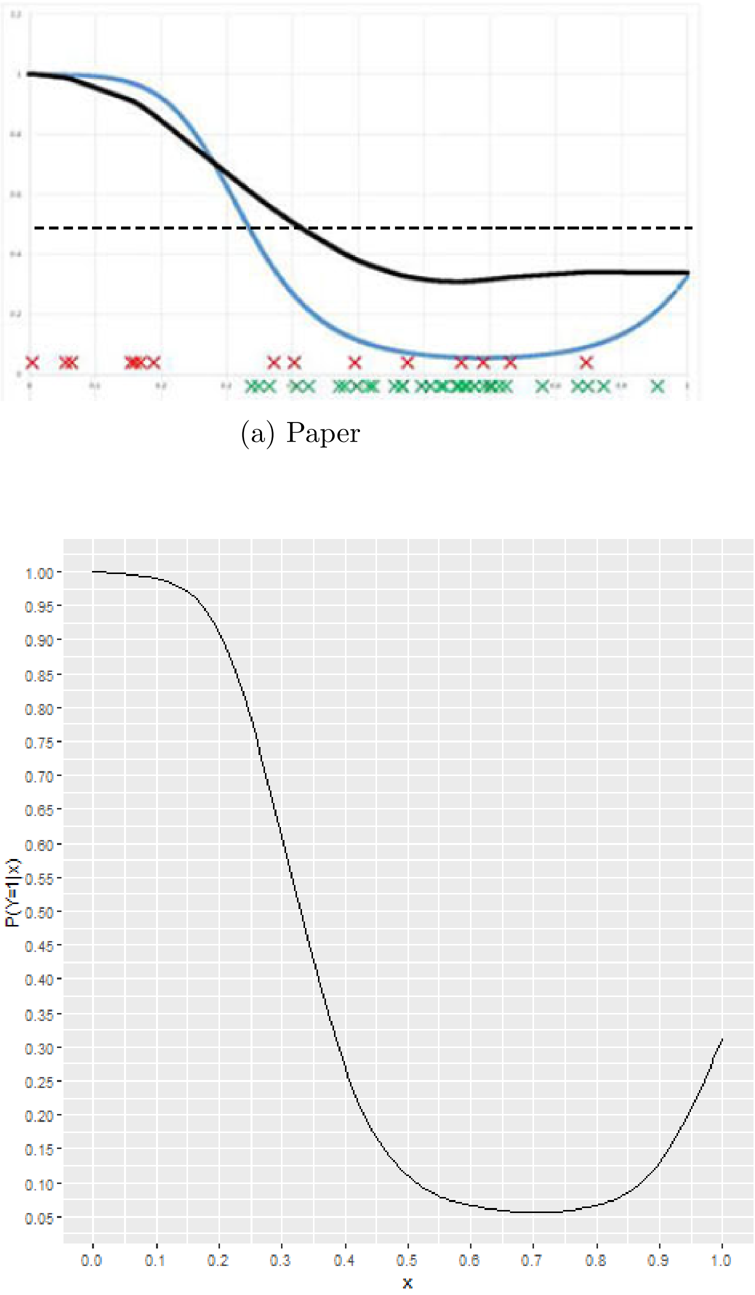


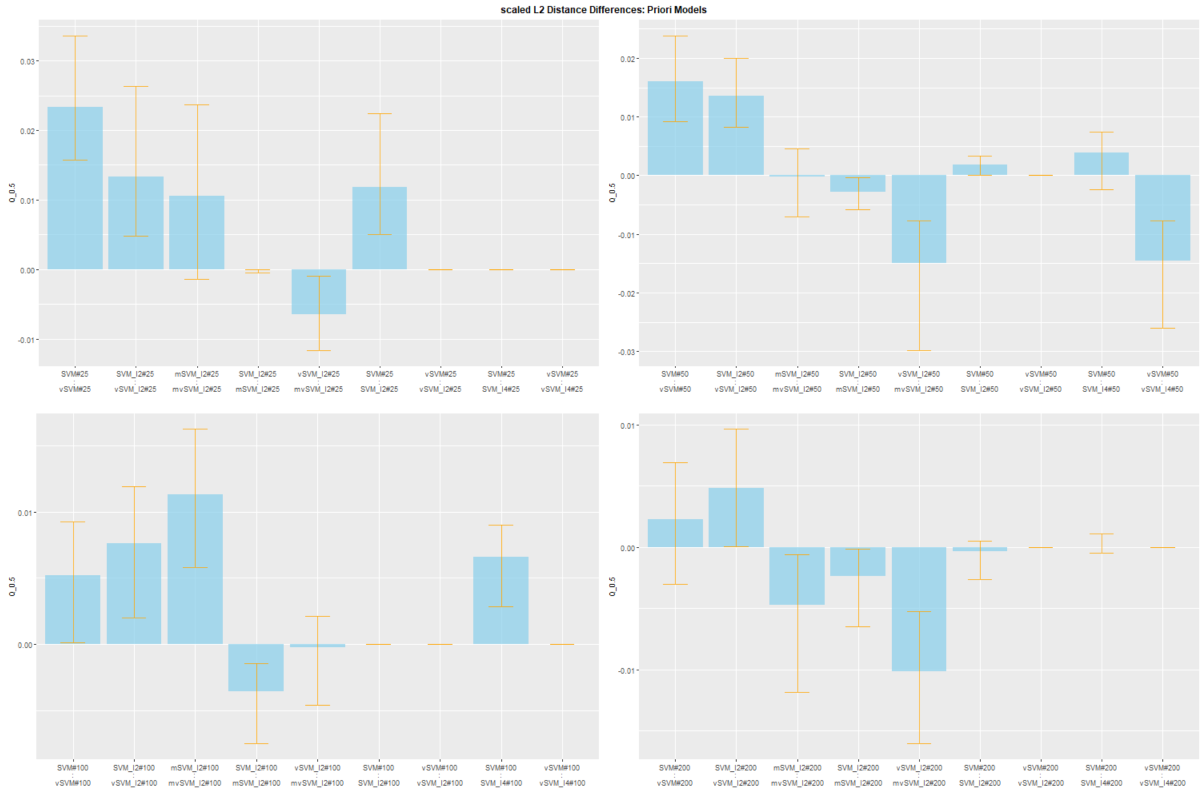
Figure B.1: Chosen beta distribution for the one dimensional space $\mathcal{X} = [0, 1]$, Beta(1.98, 2.16).



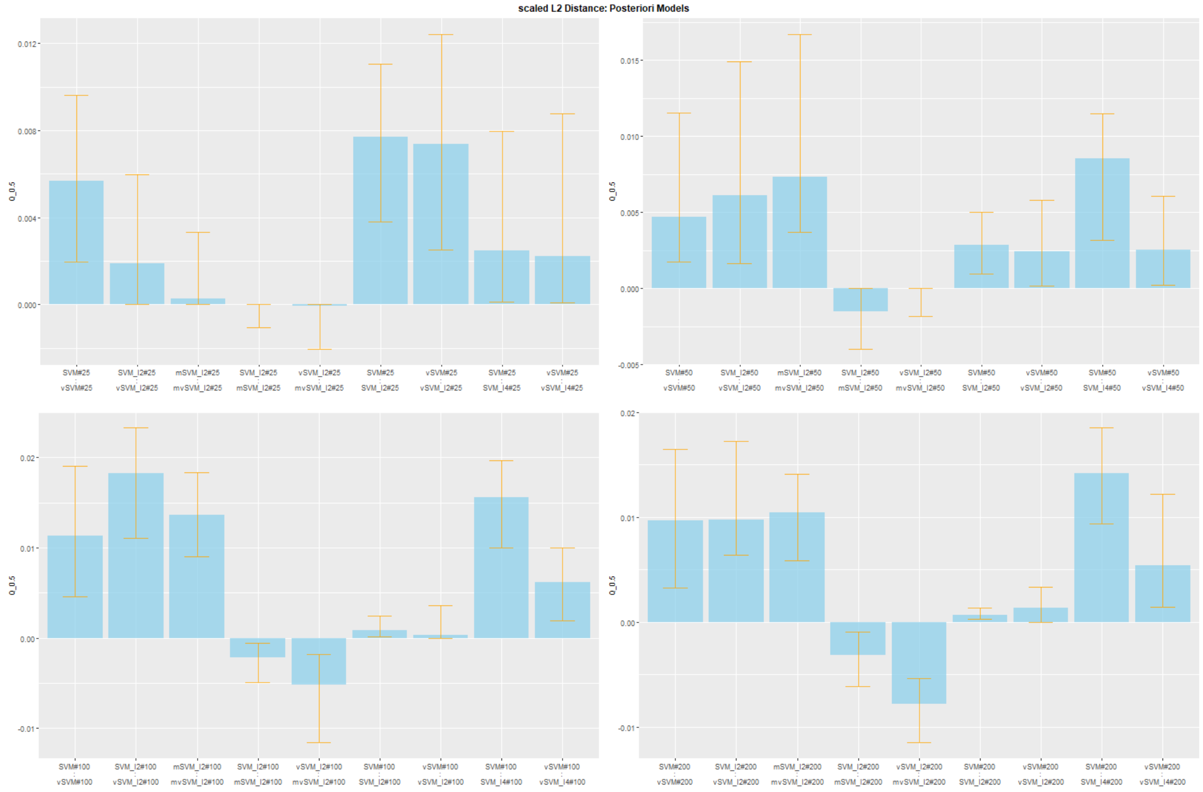
(a) Paper

(b) Reconstructed

Figure B.2: Conditional probability functions. The figure (a) is a screenshot of the original paper [11]. The blue curve is the conditional probability function applied in the paper. The figure (b) shows the reconstructed function used in this thesis.



(a) Priori



(b) Posteriori

Figure B.3: Median based evaluation of the approximation of the conditional probability function, calculated over 200 samples. (a) Priori: Median of the differences of scaled L^2 distances between the selected best models of two model classes. (b) Posteriori: Median of the differences of the smallest values of scaled L^2 distances per sample between two model classes.

Priori					
l/model	SVM : vSVM	SVM_I2 : vSVM_I2	mSVM_I2 : mvSVM_I2	SVM_I2 : mSVM_I2	vSVM_I2: mvSVM_I2
25	1.165, 6.423	0.781, 5.532	1.663, 6.222	-1.123, 4.071	2.433, 4.582
50	0.133, 7.604	-1.111,6.066	1.377, 8.094	-1.299, 5.677	0.424, 4.844
100	-0.142, 3.278	0.004, 2.961	0.894, 5.291	-1.574, 6.830	-0.426, 5.951
200	-0.251, 4.861	-0.633, 4.398	0.622, 4.611	-1.413, 6.626	-0.302, 3.987
l/model	SVM : SVM_I2	vSVM : vSVM_I2	SVM : SVM_I4	vSVM : vSVM_I4	
25	-1.099, 6.504	-2.802, 5.884	2.013, 5.596	-2.699, 4.573	
50	1.377, 5.722	-2.658, 5.774	-1.295, 3.174	0.494, 6.844	
100	-0.043, 6.167	0.507, 6.301	-1.500, 6.420	0.518, 4.415	
200	-0.343, 4.734	-1.178, 4.959	2.876, 5.350	4.444, 7.350	
Posteriori					
l/model	SVM : vSVM	SVM_I2 : vSVM_I2	mSVM_I2 : mvSVM_I2	SVM_I2 : mSVM_I2	vSVM_I2 : mvSVM_I2
25	2.281, 7.554	1.890, 6.341	1.706, 7.314	-1.095, 6.096	-1.493, 5.751
50	3.065, 4.597	1.391, 5.074	1.891, 7.773	-2.322, 5.918	-0.932, 3.096
100	0.113, 4.998	0.932, 4.564	0.651, 5.090	0.302, 7.509	-0.629, 4.677
200	-0.604, 3.462	0.408, 2.302	0.079, 4.581	-0.975, 6.328	-0.623, 3.812
l/model	SVM : SVM_I2	vSVM : vSVM_I2	SVM : SVM_I4	vSVM : vSVM_I4	
25	2.968, 7.017	1.493, 7.010	2.771, 6.909	1.146, 5.177	
50	2.107, 8.321	0.5322, 4.016	1.931, 7.088	-0.065, 4.397	
100	2.865, 4.081	1.944, 6.264	0.834, 2.987	1.610, 5.421	
200	2.159, 7.421	2.270, 5.113	0.829, 2.997	1.626, 5.035	

Table B.1: Empirical skewness (first value) and kurtosis (second value) of the differences in the pairwise evaluation regarding the L^2 distances (related to fig. 4.3).

Function approximation						
l/model	SVM_I2	vSVM_I2	mSVM_I2	mvSVM_I2	SVM_I4	vSVM_I4
25	0.05	0.4	0.05	0.05	0.1	0.4
50	0	0.4	0	0.2	0.8	0.4
100	0.05	0.4	0	0.2	0.05	0.2
200	0.4	0.4	0.2	0.2	0.05	0.4
Classification						
l/model	SVM_I2	vSVM_I2	mSVM_I2	mvSVM_I2	SVM_I4	vSVM_I4
25	0	0.4	0	0	0.1	0.4
50	0.2	0.8	0.05	0.05	0.2	0.8
100	0.4	0.4	0.1	0	0.4	0.2
200	0.8	0.05	0.1	0.1	0	0

Table B.2: The delta values of the best models according to the priori evaluation of the model classes considering invariants.