Kernel-Based Expectile Regression

Von der Fakultät Mathematik und Physik der Universität Stuttgart
zur Erlangung der Würde eines
Doktors der Naturwissenschaften (Dr. rer. nat.)
genehmigte Abhandlung

Vorgelegt von

Muhammad Farooq

aus Lahore, Pakistan

Hauptberichter:	Prof. Dr. Ingo Steinwart
Mitberichter:	Prof. Dr. Andreas Christmann
Tag der mündlichen Prüfung:	13. Oktober 2017

Institut für Stochastik und Anwendungen
der Universität Stuttgart

2017

Erklärung

Hiermit erkläre ich, dass ich die eingereichte Dissertation mit dem Titel

Kernel-Based Expectile Regression

selbständig angefertigt und keine anderen als die angegebenen Hilfsmittel benutzt habe. Alle Stellen, die dem Wortlaut oder dem Sinn nach anderen Werken, gegebenenfalls auch elektronischen Medien entnommen sind, sind von mir durch Angabe der Quelle als Entlehnung kenntlich gemacht.

....

Ort, Datum Unterschrift

Acknowledgements

I express my deepest gratitude to my doctoral advisor Prof. Dr. Ingo Steinwart who gave me the opportunity to work with him in this research area. His continuous support and constructive discussion with him on the technical challenges during the journey of my Ph.D. always helped me to achieve my goals. He was also very supportive regarding non-technical issues. I could not have imagined having a better advisor for my Ph.D. research work. Besides my advisor, I warmly thank Prof. Dr. Andreas Christmann for his interest and willingness to be a co-examiner for my doctoral thesis.

I would like to thank all my old and new colleagues of ISA who made my stay in this institute memorable by giving me the opportunity of travelling with them and having fruitful discussions with them. Special thanks go to Dr. Philipp Thomann, Simon Fischer, Ingrid Blaschzyk and Thomas Hamm for reviewing my thesis and providing their valuable comments and suggestions in order to improve it. I would also like to thank Elke Maurer for helping me in many different ways throughout my journey.

With all my heart I express my deepest gratitude to my parents, the closest persons to whom this thesis is especially dedicated. They always let me go the way I like and encourage me to achieve my goals. Words cannot express the scarifies they have made in order to fulfill my dreams. Thank you from the bottom of my heart! I am also thankful to my sisters and brother. Their love, support, and prayers were always with me during this journey.

Contents

Abstract

Conditional expectiles are becoming an increasingly important tool in finance as well as in other areas of application such as demography when the goal is to explore the conditional distribution beyond the conditional mean. In this thesis, we consider a support vector machine (SVM) type approach with the asymmetric least squares loss for estimating conditional expectiles. Firstly, we establish learning rates for this approach that are minimax optimal modulo a logarithmic factor if Gaussian RBF kernels are used and the desired expectile is smooth in a Besov sense. It turns out that our learning rates, as a special case, improve the best known rates for kernel-based least squares regression in aforementioned scenario. As key ingredients of our statistical analysis, we establish a general calibration inequality for the asymmetric least squares loss, a corresponding variance bound as well as an improved entropy number bound for Gaussian RBF kernels. Furthermore, we establish optimal learning rates in the case of a generic kernel under the assumption that the target function is in a real interpolation space.

Secondly, we complement the theoretical results of our SVM approach with the empirical findings. For this purpose we use a sequential minimal optimization method and design an SVM-like solver for expectile regression considering Gaussian RBF kernels. We conduct various experiments in order to investigate the behavior of the designed solver with respect to different combinations of initialization strategies, working set selection strategies, stopping criteria and number of nearest neighbors, and then look for the best combination of them. We further compare the results of our solver to the recent R-package ER-Boost and find that our solver exhibits a better test performance. In terms of training time, our solver is found to be more sensitive to the training set size and less sensitive to the dimensions of the data set, whereas, ER-Boost behaves the other way around. In addition, our solver is found to be faster than a similarly implemented solver for the quantile regression. Finally, we show the convergence of our designed solver.

Kurzfassung

Die *Expectile* Regression gewinnt im Finanzwesen, der *Bevölkerungswissenschaft* sowie in allen Anwendungsgebieten an Bedeutung, in denen detaillierten Eigenschaften als der Erwartungswert der bedingten Verteilung eine Rolle spielen. In dieser Arbeit betrachten wir einen Support Vector Machine (SVM) Ansatz unter Verwendung der asymmetrischen Least-Squares Verlustfunktion zur Schätzung der *Expectiles* einer bedingten Verteilung. Im ersten Abschnitt beweisen wir Lernraten für dieses Verfahren, welche bis auf einen logarithmischen Faktor minimax-optimal sind, falls wir den Gauß-Kern verwenden und die zu schätzende Expectil-Funktion in einem gewissen Besov-Raum liegt. Als Spezialfall enthält unsere Untersuchung die Least-Squares Regression und in diesem Fall liefert unser Beweis die bisher besten Raten für Gauß-Kerne. Unser Beweis stützt sich in erster Linie auf eine allgemeine Kalibrierungsungleichung der asymmetrischen Least-Squares Verlustfunktion, einer zugehörigen Varianz-Schranke, sowie einer verbesserten Schranke an die Entropie-Zahlen des Gauß-Kerns. Desweiteren beweisen wir optimale Lernraten für beliebige Kerne unter der Annahme, dass die Expectil-Funktion in einem reellen Interpolationsraum liegt.

Im zweiten Abschnitt untermauern wir die theoretischen Resultate bzgl. unseres SVM-Ansatzes mit empirischen Untersuchungen. Dazu nutzen wir eine *minimale sequentielle Optimierungsmethode*, um einen Algorithmus zur Expectil-Regression bzgl. des Gauß-Kerns zu entwickeln. Wir führen mehrere Expermimente durch, um das Verhalten unseres Algorithmuses unter verschiedenen Kombinationen von Initialisierungsstrategien, Auswahlstrategien, Stoppkriterien und Anzahlen an Nearest Neighbors zu untersuchen. Ferner führen wir einen Vergleich zwischen unserem Algorithmus und dem *R*-package ER-Boost durch, in dem feststellen, dass unser Verfahren einen geringeren Testfehler aufweist. Die Trainingszeit unseres Algorithmuses hängt stark von der Größe des Trainingsdatensatzes ab, jedoch spielt die Dimension der Daten nur eine untergeordnete Rolle. Im Gegensatz dazu verhält sich die Zeitkomplexität des ER-Boost genau umgekehrt. Zusätzlich scheint es, dass unser Algorithmus schneller ist, als ein vergleichbarer Algorithmus zur Quantil-Regression. Abschließend beweisen wir die Konvergenz unseres Optimierungsverfahrens.

Publications

Some parts of this thesis have been published in:

- **Learning Rates for Kernel-Based Expectile Regression**
  M. Farooq and I. Steinwart, Tech. Rep. 2017-003, Fakultät für Mathematik und Physik, Universität Stuttgart, 2017. `https://arxiv.org/abs/1702.07552`

- **An SVM-like Approach for Expectile Regression**
  M. Farooq and I. Steinwart, *Comput. Stat. Data Anal.*, 109:159-181, 2017. `https://doi.org/10.1016/j.csda.2016.11.010`

In addition, the source code of the **solver** for expectile regression (ex-svm) is now a part of *liquidSVM: A Fast and Versatile SVM Package*. The package can be downloaded from `http://www.isa.uni-stuttgart.de/software/`

Abbreviations

ALS	asymmetric least squares loss function
ALAD	asymmetric least absolute deviation loss function
SVM	support vector machine
TV-SVM	training-validation support vector machine
RKHS	reproducing kernel Hilbert space
RBF	radial basis function
SMO	sequential minimal optimization
i.i.d.	independently and identically distributed

Chapter 1

Introduction

Suppose that we have an input/output data set $D := ((x_1, y_1), \ldots, (x_n, y_n)) \in (X \times \mathbb{R})^n$ drawn in an i.i.d. fashion from some unknown probability distribution P on $X \times Y$, where X is an arbitrary set and $Y \subset \mathbb{R}$. In addition, suppose that there exists a probabilistic relationship between input and output variables, that is, an input value x is drawn from the marginal distribution P_X of P on X and the corresponding output value y is drawn from the conditional distribution $\mathrm{P}(Y|x), x \in X$. Then, one of the goals of statistical learning is to estimate the characteristics of the conditional distribution $\mathrm{P}(\,\cdot\,|x)$. For instance, one may be interested in estimating the central location measures of $\mathrm{P}(\,\cdot\,|x)$, that one can estimate either by the conditional mean $\mathbb{E}(\,\cdot\,|x), x \in X$ using (non)parametric least squares regression or the conditional median $\mathrm{med}(\,\cdot\,|x), x \in X$ with the help of (non)parametric median regression based on the least absolute deviation loss function. However, estimation of $\mathbb{E}(\cdot|x)$ or $\mathrm{med}(\cdot|x)$ restricts us only to the central locations' measures of the conditional distribution. In some real life applications, it is required to explore the conditional distribution $\mathrm{P}(\,\cdot\,|x)$ beyond the center of the distribution. A wonderful remark in this regard is given by Mosteller and Tukey (1977):

> *"What regression curve does is give a grand summary for the average of the distribution corresponding to the set of xs. We could go further and compute several different regression curves corresponding to the various percentage points of the distributions and thus get a more complete picture of the set."*

One effort in this direction is made by Koenker and Bassett Jr (1978) by introducing the well-known quantile regression, a natural extension of the median regression, where the conditional quantiles of $\mathrm{P}(\,\cdot\,|x)$ are obtained by minimizing the asymmetric least absolute deviation (ALAD) loss function. To be more precise, if $\mathrm{P}(\,\cdot\,|x)$ has a strictly positive Lebesgue density, then the

τ-quantile q_τ, $\tau \in (0,1)$ of Y given $x \in X$ is a solution to

$$P(Y \leq q_\tau) = \tau\,. \tag{1.1}$$

For a detailed description, different estimation methods, and the theoretical analysis for the (conditional) quantile regression, we refer the reader to Koenker (2005), Takeuchi et al (2006), Christmann and Steinwart (2007), Steinwart and Christmann (2011), and references therein.

Another approach to characterize the conditional distribution is the *expectile regression* proposed by Newey and Powell (1987). Let us denote by $Q := \mathrm{P}(Y|x)$ the conditional distribution of Y given $x \in X$. In addition we assume that the first moment of Q is finite, that is, $|Q|_1 := \int_Y y\,dQ(y) < \infty$. Then the τ-expectile $\mu_\tau^* := \mu_{\tau,Q}^* \in \mathbb{R}$ for each $\tau \in (0,1)$ is the unique solution of

$$\tau \int_{\mu_\tau^*}^{\infty} (y - \mu_\tau^*)\,dQ(y) = (1-\tau) \int_{-\infty}^{\mu_\tau^*} (\mu_\tau^* - y)\,dQ(y)\,, \tag{1.2}$$

which is also strictly monotonically increasing for $\tau \in (0,1)$, and continuously differentiable if the density of Q is continuously differentiable, see Newey and Powell (1987, Theorem 1). Unlike the quantiles that are determined by tail probabilities of Q, the expectiles are determined by the tail expectations.

One can estimate expectiles algorithmically by minimizing the expectation of a suitable loss function. There exists a class of loss functions that are consistent for expectiles. A general form of such class of loss functions can be found in Gneiting (2011, Theorem 10), see also Steinwart et al (2014, Equation (26)). However the only known convex loss function is the *asymmetric least squares loss* (ALS) that has been considered in the literature extensively. For $t \in \mathbb{R}$ and $\tau \in (0,1)$, the ALS loss is defined by

$$L_\tau(y,t) = \begin{cases} (1-\tau)(y-t)^2\,, & \text{if} \quad y < t\,, \\ \tau(y-t)^2\,, & \text{if} \quad y \geqslant t\,. \end{cases} \tag{1.3}$$

Now using (1.3), the expectile μ_τ^*, $\tau \in (0,1)$ of the distribution Q, provided that the second moment of Q is finite, can be obtained by the optimization problem of the form

$$\mu_\tau^* := \arg\min_{t \in \mathbb{R}} \mathbb{E}_{y \sim Q} L_\tau(y,t)\,, \tag{1.4}$$

see also e.g. Efron (1991) and Abdous and Remillard (1995) for further details.

Both quantiles and expectiles are special cases of so called M-quantiles as described by Breckling and Chambers (1988) and there exists a one-to-one mapping between them, see

e.g. Jones (1994) and Yao and Tong (1996). However, in general, expectiles do not coincide with quantiles of the same distribution. For instance, Jones (1994) showed that expectiles are in fact quantiles of some distribution related to the distribution of expectiles. Therefore, the choice between expectiles and quantiles mainly depends on the applications at hand, as it is the case in the duality between the mean and the median. For example, if the goal is to estimate a (conditional) threshold for which only a τ-fraction of (conditional) observations lie below this threshold, then a (conditional) τ-quantile is the right choice. On the other hand, if one is interested to estimate a (conditional) threshold for which the average distance of observation above this threshold is equal to the k times the average distance of observations below this threshold, then a τ-expectile regression is a preferable choice with $k = \frac{1-\tau}{\tau}$. Clearly, the focus in quantiles is ordering of the observations while expectiles account magnitude of the observations which makes expectiles sensitive to the extreme values of the distribution. Since expectiles estimation is computationally more efficient than quantiles estimation, one can use expectiles as a promising surrogate of quantiles in the situation where one is only interested in exploring the conditional distribution.

Expectiles have attracted considerable attention in recent years and have been applied successfully in many areas, for instance, in demography (see, Schnabel and Eilers, 2009a), in education (see, Sobotka et al, 2013) and extensively in finance, see for instance Wang et al (2011), Hamidi et al (2014), Xu et al (2016) and Kim and Lee (2016). In fact, it has recently been shown (see, e.g. Bellini et al, 2014; Steinwart et al, 2014) that expectiles are the only risk measures that enjoy the properties of coherence and elicitability, and therefore they have been suggested as potentially better alternative to the Value at Risk (VaR), see e.g. Taylor (2008), Ziegel (2016) and Bellini et al (2014). More importantly, for any $\tau \in (0,1)$, expectile immediately gives the realization of the *gain-loss ratio* or the Ω-*ratio* which is a well-known performance measure in portfolio management, see e.g. Keating and Shadwick (2002). For more applications of expectiles, we refer the interested readers to, e.g. Aragon et al (2005), Stahlschmidt et al (2014) and Guler et al (2014).

Recall (1.4) that the τ-expectiles can be computed with the help of asymmetric risks. To be more precise, for a measurable function $f : X \to \mathbb{R}$, the L-risk is defined by

$$\mathcal{R}_{L_\tau,\mathrm{P}}(f) := \int_{X\times Y} L_\tau(y, f(x))d\mathrm{P}(x,y) = \int_X \int_Y L_\tau(y, f(x))d\mathrm{P}(y|x)d\mathrm{P}_X(x)\,. \tag{1.5}$$

Then there exists a P_X-almost surely unique function $f^*_{L_\tau,\mathrm{P}} : X \to \mathbb{R}$ such that

$$\mathcal{R}_{L_\tau,\mathrm{P}}(f^*_{L_\tau,\mathrm{P}}) = \mathcal{R}^*_{L_\tau,\mathrm{P}} := \inf\{\mathcal{R}_{L_\tau,\mathrm{P}}(f) \,|\, f : X \to \mathbb{R} \text{ measurable}\}\,,$$

provided that the second moment of P is finite, that is, $|\mathrm{P}|_2 := \left(\int_{X \times Y} y^2 \, d\mathrm{P}(x,y) \right)^{1/2} < \infty$. Here, $f^*_{L_\tau,\mathrm{P}}(x)$ is the optimal decision function that is often called the Bayes decision function, and equals the τ-expectile of the conditional distribution $\mathrm{P}(\,\cdot\,|x)$ for P_X-almost all $x \in X$, that is $f^*_{L_\tau,\mathrm{P}}(x) = \mu^*_{\tau,\mathrm{P}(\cdot|x)}(x)$. A corresponding empirical estimator of $f^*_{L_\tau,\mathrm{P}}$ is denoted by $f_\mathrm{D} : X \to \mathbb{R}$ and can be obtained, for example, with the help of empirical L_τ-risks

$$\mathcal{R}_{L_\tau,\mathrm{D}}(f) = \frac{1}{n} \sum_{i=1}^{n} L_\tau(y_i, f(x_i)) \,. \tag{1.6}$$

To obtain the empirical decision function f_D, some semi-parametric and non-parametric methods have already been proposed in the literature . For instance, Schnabel and Eilers (2009b) considered penalized splines to compute smooth expectile estimates, Sobotka and Kneib (2012) proposed a couple of different procedures including least asymmetrically weighted squares in combination with mixed models, boosting within an empirical risk minimization framework, and a restricted expectiles regression model. Furthermore, a kernel method based on local linear fits was considered by Yao and Tong (1996), and a boosting method using regression trees was proposed by Yang and Zou (2015).

Another class of non-parametric estimation methods, that we consider in this work, are the so-called kernel based regularized empirical risk minimizers, which include the well known *support vector machines* (SVMs), see Vapnik (2000, p. 138ff). Recall that SVMs build a predictor $f_{\mathrm{D},\lambda}$ by solving an optimization problem of the form

$$f_{\mathrm{D},\lambda} = \arg\min_{f \in H} \lambda \|f\|_H^2 + \mathcal{R}_{L,\mathrm{D}}(f) \,. \tag{1.7}$$

Here, $\lambda > 0$ is a regularization parameter, H is a reproducing kernel Hilbert space (RKHS) over X with bounded, measurable kernel k, see e.g. Aronszajn (1950). These kernel-based methods often enjoy state-of-the-art empirical performance, relatively simple implementations, and a high flexibility. Their flexibility is based on two main ingredients, namely, the reproducing kernel Hilbert space (RKHS) H and the loss function L. The RKHS can be used to adapt to the nature of the input domain X, or more precisely, enables us to use both standard $\mathbb{R}^d$-valued data and non-standard data such as strings and graphs. Moreover, due to the so-called kernel-trick, the choice of H has little to no algorithmic consequences for solving SVM optimization problems. On the other hand, the choice of L determines the learning goal. For example, the so-called hinge loss is used for classification (Hush et al, 2006; Steinwart et al, 2011), the least squares loss leads to the conditional mean regression (Wu et al, 2006; Bauer et al, 2007; Caponnetto and De Vito, 2007; Steinwart et al, 2009; Eberts and Steinwart, 2013; Tacchetti

et al, 2013), and the ALAD loss is used to estimate conditional quantiles, see for example Steinwart and Christmann (2011) and Eberts and Steinwart (2013).

Having found an empirical estimator f_{D} (or $f_{\mathrm{D},\lambda}$ in the case of (1.7)), its quality can be measured by its distance to the target function $f^*_{L_\tau,\mathrm{P}}$, e.g. in terms of $\|f_{\mathrm{D}} - f^*_{L_\tau,\mathrm{P}}\|_{L_2(P_X)}$. For estimators obtained by some empirical risk minimization scheme, however, one can hardly ever estimate this L_2-norm directly, since $f^*_{L_\tau,\mathrm{P}}$ and $\|\cdot\|_{L_2(P_X)}$ are unknown. Instead, the standard tools of statistical learning theory give bounds on the excess risk $\mathcal{R}_{L_\tau,\mathrm{P}}(f_{\mathrm{D}}) - \mathcal{R}^*_{L_\tau,\mathrm{P}}$. Therefore, our first goal in this thesis is to establish a so-called calibration inequality of L_τ that relates both quantities. To be more precise, we will show in Theorem 3.3 that

$$C_\tau^{-1}(\mathcal{R}_{L_\tau,\mathrm{P}}(f) - \mathcal{R}^*_{L_\tau,\mathrm{P}}) \leq \|f - f^*_{L_\tau,\mathrm{P}}\|^2_{L_2(P_X)} \leq c_\tau^{-1}\left(\mathcal{R}_{L_\tau,\mathrm{P}}(f) - \mathcal{R}^*_{L_\tau,\mathrm{P}}\right), \tag{1.8}$$

holds for all $f \in L_2(P_X)$ and some constants c_τ, C_τ only depending on τ. The right hand side of (1.8) provides rates for $\|f_{\mathrm{D}} - f^*_{L_\tau,\mathrm{P}}\|_{L_2(P_X)}$ as soon as we have established rates for $\mathcal{R}_{L_\tau,\mathrm{P}}(f_{\mathrm{D}}) - \mathcal{R}^*_{L_\tau,\mathrm{P}}$. Furthermore, it is common knowledge in statistical learning theory that bounds on $\mathcal{R}_{L_\tau,\mathrm{P}}(f_{\mathrm{D}}) - \mathcal{R}^*_{L_\tau,\mathrm{P}}$ can be improved if so-called variance bounds are available. We will see in Lemma 3.4 that the right hand side of (1.8) leads to an optimal variance bound for L_τ whenever Y is bounded. Note that both (1.8) and the variance bound are independent of the considered expectile estimation method. In fact, both results are key ingredients for the statistical analysis of any expectile estimation method based on some form of empirical risk minimization. In addition, we will show in Lemma 4.12 that (1.8) leads to establish a bound for approximation error function in the case of generic kernels if the target function is in a real interpolation space, that is, $f^*_{L_\tau,\mathrm{P}} \in [L_2(P_X), H]_{\beta,\infty}$ for some $\beta \in (0,1)$, where H is a RKHS for a generic kernel.

Our second goal is to establish learning rates of the SVM-type algorithm (1.7). Since $2L_{1/2}$ equals the least squares loss, any statistical analysis of (1.7) also provides results for SVMs using the least squares loss. The latter have already been extensively investigated in the literature. For example, learning rates for generic kernels can be found in Cucker and Smale (2002), De Vito et al (2005), Caponnetto and De Vito (2007), Steinwart et al (2009), Mendelson and Neeman (2010) and references therein. Among these articles, only Cucker and Smale (2002), Steinwart et al (2009) and Mendelson and Neeman (2010) obtain learning rates in minimax sense under some specific assumptions. For example, Cucker and Smale (2002) assume that the target function $f^*_{L_{1/2},\mathrm{P}} \in H$, while Steinwart et al (2009) and Mendelson and Neeman (2010) establish optimal learning rates for the case in which H does not contain the target function.

Recently Eberts and Steinwart (2013) have established (essentially) asymptotically optimal learning rates for least squares SVMs using Gaussian RBF kernels under the assumption that the target function $f^*_{L_{1/2},\mathrm{P}}$ is contained in some Sobolev or Besov space. Recall that the Gaussian RBF kernels are defined by

$$k_\gamma(x,x') := \exp(-\gamma^{-2}\|x-x'\|_2^2)\,, \qquad x,x' \in \mathbb{R}^d\,,$$

where $\gamma > 0$ is called the width parameter that is usually determined in a data-dependent way, e.g. by cross-validation. A key ingredient of the work of Eberts and Steinwart (2013) is to control the capacity of RKHS $H_\gamma(X)$ for Gaussian RBF kernel k_γ on the closed unit Euclidean ball $X \subset \mathbb{R}^d$ by an entropy number bound

$$e_i(\mathrm{id} : H_\gamma(X) \to l_\infty(X)) \le c_{p,d}(X)\,\gamma^{-\frac{d}{p}}\,i^{-\frac{1}{p}}, \qquad i \ge 1\,,$$

see Steinwart and Christmann (2008, Theorem 6.27), which holds for all $\gamma \in (0,1]$ and all $p \in (0,1]$. Unfortunately, the constant $c_{p,d}(X)$ derived from Steinwart and Christmann (2008, Theorem 6.27) depends on p in an unknown manner. As a consequence, Eberts and Steinwart (2013) were only able to show learning rates of the form

$$n^{-\frac{2\alpha}{2\alpha+d}+\xi}$$

for all $\xi > 0$. To address this issue, we use (van der Vaart and van Zanten, 2009, Lemma 4.5) and derive the following improved entropy number bound

$$e_i(\mathrm{id} : H_\gamma(X) \to l_\infty(X)) \le (3K)^{\frac{1}{p}} \left(\frac{d+1}{ep}\right)^{\frac{d+1}{p}} \gamma^{-\frac{d}{p}} i^{-\frac{1}{p}}, \qquad i \ge 1\,, \tag{1.9}$$

which holds for all $p \in (0,1]$ and $\gamma \in (0,1]$ and some constant K only depending on d. Note that (1.9) provides an upper bound for $c_{p,d}(X)$ whose dependence on p is explicitly known. Using this new bound, we are then able to establish improved learning rates of the form

$$(\log n)^{d+1} n^{-\frac{2\alpha}{2\alpha+d}}\,. \tag{1.10}$$

Clearly these new rates replace the nuisance factor n^ξ of learning rates of Eberts and Steinwart (2013) by some logarithmic term. Up to this logarithmic factor our new rates are minimax optimal (see Györfi et al, 2002, Chapter 1.7) if $f^*_{L_\tau,\mathrm{P}} \in W_2^\alpha(\mathbb{R}^d)$ for $\alpha > \frac{d}{2}$ or if $f^*_{L_\tau,\mathrm{P}} \in B^\alpha_{2,\infty}(\mathbb{R}^d)$ for $\alpha > d$. In addition, our statistical analysis provides learning rates for all asymmetric cases, that is, for $\tau \neq 1/2$, which have not been established in the literature yet, and also were not possible to induce from the work of Eberts and Steinwart (2013).

Besides learning rates for the Gaussian RBF kernels (1.10), we also establish learning rates for generic kernels. For this we further assume that $Y \subset [-M, M]$, $M > 0$ and that the target function is in a real interpolation space, i.e. $f^*_{L_\tau,\mathrm{P}} \in [L_2(P_X), H]_{\beta,\infty}$ for some $\beta \in (0,1)$, Then we obtain optimal learning rates of the form

$$n^{-\frac{\beta}{\beta+p}},$$

where $p \in (0,1)$. For $\tau = 0.5$, these rates are the same as the ones obtained by Steinwart et al (2009) in case of the least squares loss. However, the advantage with our rates is that they hold, modulo a constant term, for all $\tau \in (0,1)$.

Our third goal in this thesis is to complement the above mentioned theoretical results of SVMs for Gaussian RBF kernels with the empirical findings. For this purpose, we design in the following an SVM-like solver for solving the optimization problem (1.7). Note that, besides Huang et al (2014) who have considered a kernelized iteratively reweighted strategy, no fully adaptive and efficient solver for the ALS loss has been proposed yet. For designing the solver, let us fix a feature space H_0 and a feature map $\Phi : X \to H_0$ of the considered kernel k. Then for all $\mathbf{x} \in X$, one can represent $f \in H$ in terms of $\mathbf{w} \in H_0$ via

$$f(\mathbf{x}) = \langle \mathbf{w}, \phi(\mathbf{x}) \rangle_{H_0}, \tag{1.11}$$

see Steinwart and Christmann (2008, Theorem 4.21) for further details. Note that the latter theorem also shows that

$$\|f\|_H = \inf\{\|\mathbf{w}\|_{H_0} : \mathbf{w} \in H_0 \text{ with } f = \langle \mathbf{w}, \phi(\cdot)\rangle_{H_0}\}. \tag{1.12}$$

By using (1.6) together with (1.12) in (1.7), we then obtain the standard regularized problem for SVMs without offset

$$\arg\min_{\mathbf{w}\in H_0} \lambda\|\mathbf{w}\|^2_{H_0} + \frac{1}{n}\sum_{i=1}^{n} L_\tau(y_i, f(x_i)). \tag{1.13}$$

About the last two decades the SVM-algorithms *without offset* have been considered because the offset term does in general not promise any theoretical and empirical advantages if one consider large RKHSs such as Gaussian RKHSs, see e.g. Vogt (2002), Steinwart (2003), Keerthi et al (2006), Steinwart et al (2011) and references theirin. On the contrary, the offset term imposes more restrictions on the solver. We will discuss on it in more details in Chapter 5.1. Now, we

reformulate (1.13) and obtain the primal optimization problem of the form

$$\begin{aligned} &\arg\min_{\substack{(\mathbf{w},\xi_+,\xi_-)\\ \mathbf{w}\in H}} \mathcal{P}_C(\mathbf{w},\xi_+,\xi_-) := \frac{1}{2}\|\mathbf{w}\|^2 + C\tau\sum_{i=1}^n \xi_{i,+}^2 + C(1-\tau)\sum_{i=1}^n \xi_{i,-}^2\,, \\ &\text{such that} \qquad \xi_{i,+} \geq y_i - \langle \mathbf{w}, \phi(\mathbf{x}_i)\rangle\,, \\ &\qquad\qquad \xi_{i,-} \geq \langle \mathbf{w}, \phi(\mathbf{x}_i)\rangle - y_i\,, \\ &\qquad\qquad \xi_{i,+}\,, \xi_{i,-} \geq 0\,, \qquad \forall\, i = 1,\ldots,n\,, \end{aligned} \tag{1.14}$$

where $C := \frac{1}{2n\lambda} > 0$. Using standard Lagrangian techniques, one can easily obtain the dual optimization problem

$$\begin{aligned} &\arg\max_{(\alpha,\beta)} \mathcal{D}(\alpha,\beta) := \langle \alpha-\beta, \mathbf{y}\rangle - \frac{1}{2}\langle \alpha-\beta, K(\alpha-\beta)\rangle - \frac{1}{4C\tau}\langle \alpha,\alpha\rangle - \frac{1}{4C(1-\tau)}\langle \beta,\beta\rangle\,, \\ &\qquad\qquad \alpha_i \geq 0, \beta_i \geq 0\,, \qquad \forall\, i = 1,\ldots,n\,. \end{aligned} \tag{1.15}$$

Here $\mathbf{y}$ is the $n\times 1$ vector of labels and K is the $n\times n$ matrix with entries $K_{i,j} := k(x_i, x_j)$, $i,j = 1,\ldots,n$. The convexity of the loss (1.3) leads to the convexity of (1.14) which as a result leads to the concavity of (1.15). This ensures the fulfillment of the strong duality assumptions and consequently, the primal optimal solution can be obtained from a dual optimal solution using a simple transformation. To be more precise, if (α^*, β^*) is an optimal solution of the dual problem (1.15), then the optimal solution of the corresponding primal problem (1.14), for fixed $\tau \in (0,1)$, can be obtained by

$$\begin{aligned} \mathbf{w}^* &:= \sum_{i=1}^n (\alpha_i^* - \beta_i^*)\phi(\mathbf{x}_i)\,, \\ \xi_{i,+}^* &:= \max\left\{0, y_i - \langle \mathbf{w}^*, \phi(\mathbf{x}_i)\rangle\right\}, \\ \xi_{i,-}^* &:= \max\left\{0, \langle \mathbf{w}^*, \phi(\mathbf{x}_i)\rangle - y_i\right\}, \end{aligned} \tag{1.16}$$

Moreover, we obtain for $\mathcal{D}^* := \mathcal{D}(\alpha^*, \beta^*)$ and $\mathcal{P}^* := \mathcal{P}_C^*(\mathbf{w}^*, \xi_+^*, \xi_-^*)$ that $\mathcal{D}^* = \mathcal{P}^*$. We further obtain for fixed $\tau \in (0,1)$ the following dual representation of the empirical conditional expectile estimator $f_{\mathrm{D},\lambda}$ defined by (1.7)

$$f_{\mathrm{D},\lambda}(\cdot) = \sum_{i=1}^n (\alpha_i^* - \beta_i^*) k(\,\cdot\,, x_i)\,. \tag{1.17}$$

In order to achieve (1.17), we need to solve (1.15). Since (1.15) is of quadratic nature, we can implement quadratic programming (QP) techniques to solve (1.15). In the literature, we find many quadratic techniques for this purpose. However, in this thesis, we use the limiting case of decomposition method, namely, the Sequential Minimal Optimization (SMO) method

that optimizes two coordinates at each iteration (Platt, 1999). Note that the without-offset version of SVM allows us to design an SMO-type algorithm, namely, the 1D algorithm that can update one coordinate per iteration. For constants $b_1, b_2 \in (1, \infty)$ and $c_i \in \mathbb{R}$, $i \in \{1, \ldots, n\}$, we will show in Theorem 5.2 that this algorithm finds the 1D-feasible solution of (1.15) by using

$$\alpha_i^+ = \max\left(0, \frac{c_i}{b_1}\right), \qquad \beta_i^+ = \max\left(0, -\frac{c_i}{b_2}\right),$$

where $c_i := y_i - \sum_{j \neq i=1}^{n} (\alpha_j - \beta_j) k(x_i, x_j)$ and then chooses the best direction i^* using the 1D-gain of the dual objective function. For constants $b_1, b_2 \in (1, \infty)$ and $\delta, \eta \in \mathbb{R}$, the 1D-gain for each $i \in \{1, \ldots, n\}$ is obtained by

$$G(\delta_i, \eta_i) := \delta_i \left(\nabla \mathcal{D}_{\alpha_i}(\alpha, \beta) - \frac{b_1 \delta_i}{2} \right) + \eta_i \left(\nabla \mathcal{D}_{\beta_i}(\alpha, \beta) - \frac{b_2 \eta_i}{2} \right) + \delta_i \eta_i \,,$$

where $\delta_i = \alpha_i^+ - \alpha_i$ and $\eta_i = \beta_i^+ - \beta_i$ denote the difference between the new and the old values of α_i and β_i respectively, and $\nabla \mathcal{D}_{\alpha_i}(\alpha, \beta)$ and $\nabla \mathcal{D}_{\beta_i}(\alpha, \beta)$ are the gradients of $\mathcal{D}(\alpha, \beta)$ w.r.t. α_i and β_i, respectively. Besides that, we establish a duality gap criterion to determine when 1D solver stops iterating. In addition, we design initialization strategies, namely, cold start and warm start, where the former initialize the solver with zeros and latter by recycling the old solution. Note that our designed SMO-type algorithm uses more than first order information, namely, the quadratic and concave nature of $\mathcal{D}(\alpha, \beta)$ and exactly maximizing the gain in the dual during each iteration. Extending the idea of the 1D algorithm, we also design an SMO algorithm that updates two dual coordinates per iteration, see Section 5.3 for further details. In order to to obtain the optimal solution for (1.15), using only either the 1D algorithm above or a 2D algorithm that looks for the best pair of directions is not a suitable choice, because the former takes a longer time to converge and the latter requires a $O(n^2)$ search, see Steinwart et al (2011) in the case of the hinge loss. We therefore design two low-cost best direction search strategies, namely, WSS 1 and WSS 2. The former searches for two 1D directions from two equal splits of the index set $\{1, \ldots, n\}$, say i^* and j^*, respectively, for which the 1D-gain is maximum, and the latter first fix the i^* chosen by WSS 1, and then searches for another direction j^* based on the maximum 2D gain from k-nearest neighbors of x_i with the metric $d(x, x') := \|x - x'\|^2$. We also show the theoretical convergence of our solver for expectile regression in Section 5.4.

The behavior of the designed solver for the expectile regression is investigated by conducting various experiments. It turns out that the solver performs at its best when one chooses the warm start initialization method, the WSS 2 working set selection strategy, the nearest neighbors size 15 and the duality gap without clipping as a stopping criterion. On the contrary, Steinwart

et al (2011) show that their solver for the hinge loss performs at its best when clipped duality gap is used as a stopping criterion with nearest neighbors size 10 while keeping the setting of aforementioned others criteria the same. We further compare the performance of our solver with respect to test error and training time to the R-package ER-Boost proposed by Yang and Zou (2015) for expectile regression. The results show, see Section 5.5, that the test performance of our solver is better than ER-Boost on various data sets. Regarding training time, we observe that our solver is more sensitive to the training set size and less sensitive to the dimensions of the data set, whereas, ER-Boost behaves the other way around. Finally, recall that one can use the expectiles as a computationally surrogate of the quantiles if one is interested only to explore the conditional distribution. We therefore compare the run times of our solver to the run time of the solver for quantile regression. It turns out the expectile solver is, depending on the data set size of the considered examples in Section 5.5, between 2 and 10 times faster than the solver for quantile regression, which shows the clear computational advantage of using expectile regression over quantile regression.

The rest of the thesis is organized as follows: Chapter 2 introduces some basic concepts, which include some properties of losses and their risks (Section 2.1), basics of kernels and their RKHSs (Section 2.2), a brief overview of the statistical analysis of SVMs (Section 2.3) and basic concepts of working with convex optimization problems (Section 2.4). In Chapter 3, we characterize the ALS loss function. Besides establishing Lipschitz continuity bounds for the ALS loss (Lemma 3.1), the so-called self-calibration inequalities (Theorem 3.3) are the main results of this chapter. These inequalities are then used to establish variance bounds in Lemma 3.4 for the ALS loss. The self-calibration inequalities and the corresponding variance bounds together with improved entropy bounds for Gaussian RKHSs (Lemma 4.2) are used as the key ingredients in Chapter 4 for establishing oracle inequalities (Theorem 4.6) and minimax optimal learning rates (Corollary 4.7) for SVMs under the assumption that $Y \subseteq [-M, M]$, $M > 0$ and the target function is smooth in a Besov sense (Section 4.1.3). In Section 4.1.4, we use a data-dependent parameter selection method that splits the data set D into a training and a validation set and achieves same learning rates adaptively, that is, without knowning the unknown smoothness parameters. Furthermore, we replace the assumption of bounded regression with the assumption of exponential decay of Y-tails in Section 4.1.5 and achieve the same learning rates. Finally, in Section 4.2, we consider generic kernels and obtain the learning rates under the assumption that $Y \subseteq [-M, M]$ and that the target function is in a real interpolation space.

In Chapter 5 we design an SVM-like solver for expectile regression. This includes the formulation of the primal and the dual optimization problem for our learning scenario (Section 5.1), an algorithm for updating one coordinate along with some initialization strategies (Section 5.2) and an algorithm for updating two coordinates with some working set selection strategies (Section 5.3). In addition, the convergence analysis of the designed solver is given in Section 5.4. Finally, experimental results are presented in Section 5.5 where we investigate the behavior of the solver and compare its performance with the performance of existing R-package ER-Boost. The detailed results of the experiments are given in the appendix A.

In the end, we would like to mention that many of the results presented in this thesis have been published in advance. For instance, the results of Chapter 3 and partly of Chapter 4 have been published in Farooq and Steinwart (2017a). Moreover, the findings of Chapter 5 have been published in Farooq and Steinwart (2017b). Furthermore, the source code of the solver for expectile regression (ex-svm) has been added in the larger package liquidSVM, see Steinwart and Thomann (2017), that can be downloaded from `http://www.isa.uni-stuttgart.de/software/`.

Chapter 2

Fundamentals

This chapter introduces some basic concepts which we use in the subsequent chapters. In Section 2.1 we present some notions of loss functions and their associated risks which will extensively be used in Chapter 3 to characterize the ALS loss function. Section 2.2 deals with basic concepts of kernels and their reproducing kernel Hilbert spaces. We also briefly describe the RKHSs of the well-known Gaussian RBF kernels that are often used in SVMs. After this, an overview of the statistical analysis of SVMs is given in Section 2.3. The concepts given in both Section 2.2 and Section 2.3 will be used in Chapter 4 in order to establish oracle inequalities and learning rates for the optimization problem (1.7). Finally, Section 2.4 covers the basics concepts of convex optimization that will be used in Chapter 5 to develop an algorithm using an SVM-like approach for solving (1.7). The contents of this chapter are primarily based on Cristianini and Shawe-Taylor (2000), Schölkopf and Smola (2002), Boyd and Vandenberghe (2004), Abe (2005) and Steinwart and Christmann (2008).

2.1 Some Properties of Losses and Their Risks

Given an i.i.d. data set $D := ((x_1, y_1), \ldots, (x_n, y_n)) \in (X \times Y)^n$ drawn from some unknown probability distribution P on $X \times Y$, where $X \subset \mathbb{R}^d$ and $Y \subset \mathbb{R}$, the goal of (supervised) statistical learning is to find a function $f : X \to \mathbb{R}$ such that for every pair $(x, y) \in (X \times Y)$, the evaluation $f(x)$ is a good prediction of the possible response y at x. In order to assess the quality of the "learned" function f, we recall some well established concepts from Steinwart and Christmann (2008, Chapter 2 and Chapter 3) and Schölkopf and Smola (2002, Chapter 3). Let us begin by introducing the notion of loss function that measures the *loss* or *cost* of predicting response y at different levels of input variable(s).

Definition 2.1 (cf. Steinwart and Christmann (2008, Definition 2.1)). *Let $(X, \mathcal{A})$ be a measurable space and $Y \subset \mathbb{R}$ be a closed subset. Then a function $L : X \times Y \times \mathbb{R} \to [0, \infty)$ is called a **loss function** if it is measurable.*

We often use the notation $L \circ f$ to represent the function $(x, y) \to L(x, y, f(x))$. The loss function L can either be a *supervised* loss function defined by $L := Y \times \mathbb{R} \to [0, \infty)$ or an *unsupervised* loss function defined by $L : X \times \mathbb{R} \to [0, \infty)$. In practice, the choice of loss function is determined by the learning problem at hand. For instance, in the case of the supervised loss, the *classification loss* $L_{\text{class}} := \{-1, 1\} \times \mathbb{R} \to [0, \infty)$ defined by $L_{\text{class}}(y, t) := \mathbf{1}_{(-\infty,0]}(y\text{sign}t)$ is used for the classification problem and the *least squares loss* $L_{\text{LS}} := Y \times \mathbb{R} \to [0, \infty)$ defined by $L_{\text{LS}}(y, t) := (y - t)^2$ is used for prediction. Furthermore, to study quantiles and expectiles, the *pin-ball loss* and the *asymmetric least squares loss* are used respectively, see Steinwart and Christmann (2011) and Farooq and Steinwart (2017b) for further details. Note that a loss function can be characterized by its desirable properties. We define in the following the convexity and continuity of the loss function, see e.g. Steinwart and Christmann (2008, Definition 2.12 and 2.14), and we will further see in Chapter 5 that how the convex loss function leads to the convex optimization problem.

Definition 2.2 (cf. Steinwart and Christmann (2008, Definition 2.12 and 2.14)). *A loss $L : X \times Y \times \mathbb{R} \to [0, \infty)$ is called **(strictly) convex** and **continuous** if $L(x, y, \cdot) : \mathbb{R} \to [0, \infty)$ is (strictly) convex and continuous, respectively, for all $x \in X$ and $y \in Y$.*

Recall Definition 2.1 that the loss function L measures only the loss of a function f for a fixed pair (x, y). In statistical learning, we are rather interested in the *average loss*, where the average is taken with respect to the probability distribution P.

Definition 2.3 (cf. Steinwart and Christmann (2008, Definition 2.2 and 2.3)). *For a loss function $L : X \times Y \times \mathbb{R} \to [0, \infty)$ and a probability distribution P on $X \times Y$, the **L-risk** of a measurable function $f : X \to \mathbb{R}$ is defined by*

$$\mathcal{R}_{L,\mathrm{P}}(f) := \int_{X \times Y} L(x, y, f(x)) \mathrm{dP}(x, y) = \int_X \int_Y L(x, y, f(x)) \mathrm{dP}(y|x) \mathrm{dP}_X(x) \,. \tag{2.1}$$

*Moreover, the **minimal L-risk** is defined by*

$$\mathcal{R}^*_{L,\mathrm{P}} := \inf\{\mathcal{R}_{L,\mathrm{P}}(f) | f : X \to \mathbb{R} \text{ is measurable}\} \,,$$

*which is also called the **Bayes risk** for some loss function L with respect to P.*

Here, the integral over $X \times Y$ always exists because $(x, y) \mapsto L(x, y, f(x))$ is measurable and non-negative. If there exists a measurable function $f^*_{L,\mathrm{P}} : X \to \mathbb{R}$ such that $\mathcal{R}_{L,\mathrm{P}}(f^*_{L,\mathrm{P}}) = \mathcal{R}^*_{L,\mathrm{P}}$, then $f^*_{L,\mathrm{P}}$ is called the *Bayes decision function* (we will often call an optimal decision function). We will see in Chapter 3 in the case of the ALS loss that $f^*_{L,\mathrm{P}}$ is unique, however, in some other cases it is not, see e.g. Steinwart and Christmann (2011) for the pinball loss.

The risk function $\mathcal{R}_{L,\mathrm{P}}(\cdot)$ is measurable in the following scenario, see Steinwart and Christmann (2008, Lemma 2.11) for proof. Assume that $\mathcal{F} \subset \mathcal{L}_0(X) := \{f : X \to \mathbb{R} | f \text{ measurable}\}$ is a subset equipped with a complete and separable metric d, and the corresponding Borel σ-algebra. We also assume that

$$\lim_{n\to\infty} d(f_n, f) = 0 \qquad \Longrightarrow \qquad \lim_{n\to\infty} f_n(x) = f(x)\,, \qquad x \in X\,,$$

for all $f_n, f \in \mathcal{F}$, that is d dominates the pointwise convergence. Then the map $\mathcal{F} \times X \to \mathbb{R}$, $(f, x) \mapsto f(x)$ is measurable and thus are the map $X \times Y \times \mathcal{F} \to [0, \infty)$, $(x, y, f) \mapsto L(x, y, f(x))$ and the risk functional $\mathcal{R}_{L,P} : \mathcal{F} \to [0, \infty]$. Here, it is interesting to note that the pointwise convergence of the sequence of measurable functions (f_n) to some $f : X \to \mathbb{R}$ implies the convergence of $L(x, y, f_n(x)) \to L(x, y, f(x))$ for all $(x, y) \in X \times Y$. However, this does not generally hold for the convergence of associated risk $\mathcal{R}_{L,\mathrm{P}}(f_n)$ to $\mathcal{R}_{L,\mathrm{P}}(f)$. In other words, the risk of a continuous loss is not necessarily continuous. In that case, one can measure (local) Lipschitz continuity that holds for almost all frequently used loss functions.

Definition 2.4 (cf Steinwart and Christmann (2008, Definition 2.18))**.** *A loss function* $L : X \times Y \times \mathbb{R} \to [0, \infty)$ *is called*

i) ***locally Lipschitz continuous*** *if for all* $M > 0$*, we have*

$$|L|_{M,1} := \sup_{\substack{t,t' \in [-M,M] \\ t \neq t'}} \sup_{\substack{x \in X \\ y \in Y}} \frac{L(x, y, t) - L(x, y, t')}{|t - t'|} < \infty\,.$$

ii) ***Lipschitz continuous*** *if* $|L|_1 := \sup_{M>0} |L|_{M,1} < \infty$.

If $Y \subset \mathbb{R}$ is finite and $L : Y \times \mathbb{R} \to [0, \infty)$ is a convex loss function, then by Steinwart and Christmann (2008, Lemma A.6.5), the loss L is locally Lipschitz continuous, and by Steinwart and Christmann (2008, Lemma 2.13 and Lemma 2.19), $\mathcal{R}_{L,\mathrm{P}} : \mathcal{L}_0(X) \to [0, \infty]$ is convex and locally Lipschitz continuous, respectively. To be more precise, for all $M > 0$ and all $f, g \in L_\infty(P_X)$ with $\|f\|_\infty \leq M$ and $\|g\|_\infty \leq M$, we have, see Steinwart and Christmann (2008, Lemma 2.19)

$$|\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}_{L,\mathrm{P}}(g)| \leq |L|_{M,1} \cdot \|f - g\|_{L_1(P_X)}\,.$$

In the following, we present the notion of Nemitski loss.

Definition 2.5 (cf. Steinwart and Christmann (2008, Definition 2.16)). *A loss $L : X \times Y \times \mathbb{R} \to [0, \infty)$ is called a* ***Nemitski loss*** *if there exists a measurable function $b : X \times Y \to [0, \infty)$ and an increasing function $h : [0, \infty) \to [0, \infty)$ such that*

$$L(x, y, t) \leq b(x, y) + h(|t|)\,, \qquad (x, y, t) \in X \times Y \times \mathbb{R}\,.$$

Furthermore, L is called a ***Nemitski loss of order*** *$p \in (0, \infty)$, if there exists a constant $c > 0$ such that*

$$L(x, y, t) \leq b(x, y) + c\,|t|^p\,, \qquad (x, y, t) \in X \times Y \times \mathbb{R}\,.$$

*Finally, if P is a distribution on $X \times Y$ with $b \in \mathcal{L}_1(\mathrm{P})$, we say L is P-**integrable Nemitski loss**.*

From Definition 2.3 it is trivial to see that the risk function (2.1) can be computed by iterated integrals. In other words, one can compute inner and outer integrals of (2.1) separately. This generates the idea of *inner risks* that are key ingredients of our analysis in Chapter 3.

Definition 2.6 (cf. Steinwart and Christmann (2008, Definition 3.3)). *For a loss $L : X \times Y \times \mathbb{R} \to [0, \infty)$ and a distribution Q on Y, the* ***inner L-risk*** *of Q are defined by*

$$\mathcal{C}_{L,Q,x}(t) := \int_Y L(x, y, t) dQ(y)\,, \qquad x \in X,\ t \in \mathbb{R}\,,$$

and the ***minimal inner L-risks*** *are defined by*

$$\mathcal{C}^*_{L,Q,x} := \inf_{t \in \mathbb{R}} \mathcal{C}_{L,Q,x}(t)\,, \qquad x \in X\,.$$

Given a distribution P on $X \times Y$, the inner risks $\mathcal{C}_{L,P(\cdot|x),x}(f)$ of a function f can be used to compute the risk $\mathcal{R}_{L,\mathrm{P}}(f)$ by

$$\mathcal{R}_{L,\mathrm{P}}(f) = \int_X \mathcal{C}_{L,\mathrm{P}(\cdot|x),x}(f(x)) \mathrm{dP}_X(x)\,.$$

Furthermore, Steinwart and Christmann (2008, Lemma 3.4 and Lemma 3.11) show that the minimal inner risk $\mathcal{C}^*_{L,P(\cdot|x),x}$ is measurable in $x \in X$ and finite. Therefore, $\mathcal{R}^*_{L,\mathrm{P}}$ can be computed by

$$\mathcal{R}^*_{L,\mathrm{P}} = \int_X \mathcal{C}^*_{L,\mathrm{P}(\cdot|x),x} \mathrm{dP}_X(x)\,.$$

In other words, the minimal risk $\mathcal{R}^*_{L,\mathrm{P}}$ can be achieved by *pointwisely minimizing* the inner risks $\mathcal{C}_{L,\mathrm{P}(\cdot|x),x}$, $x \in X$, which, in general, is easier than direct minimization of $\mathcal{R}_{L,\mathrm{P}}(\cdot)$. Moreover, one can compute the *excess L-risk*, when $\mathcal{R}^*_{L,\mathrm{P}} < \infty$ holds, by

$$\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}} = \int_X \mathcal{C}_{L,\mathrm{P}(\cdot|x),x}(f(x)) - \mathcal{C}^*_{L,\mathrm{P}(\cdot|x),x} \mathrm{dP}_X(x)\,, \tag{2.2}$$

for all measurable $f : X \to \mathbb{R}$. Clearly, one can obtain the excess risk first by analyzing the excess inner L-risks $\mathcal{C}_{L,\mathrm{P}(\cdot|x),x}(f(x)) - \mathcal{C}^*_{L,\mathrm{P}(\cdot|x),x}, x \in X$ and then investigating the integration with respect to P_X, see Steinwart (2007). Besides some technical advantages, the analysis only depends on P via the conditional distributions $\mathrm{P}(\cdot|x)$ and hence allows us to consider the excess inner L-risks $\mathcal{C}_{L,Q,x}(f(x)) - \mathcal{C}^*_{L,Q,x}$ for classes of distributions Q on Y as a *template* for $\mathcal{C}_{L,\mathrm{P}(\cdot|x),x}(f(x)) - \mathcal{C}^*_{L,\mathrm{P}(\cdot|x),x}$. This idea is very useful in the context of machine learning where we assume that the distribution P and hence $\mathrm{P}(\cdot|x), x \in X$, is (almost) completely unknown, and the only information we have is that the distribution P belongs to a group of a certain type of distributions.

We conclude this section by presenting the idea of clipping that was first used by Bousquet and Elisseeff (2002) in the context of SVMs. Here we assume that $Y \subset [-M, M]$ for some $M > 0$, and we are interested in $[-M, M]$-valued estimator on X. For this, we need to restrict the loss L to $X \times Y \times [-M, M]$.

Definition 2.7 (cf. Steinwart and Christmann (2008, Definition 2.22)). *Let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a loss function and $M > 0$. Then we say that L can be **clipped** at M, if for all $(x, y, t) \in (X \times Y \times \mathbb{R})$ we have*

$$L(x, y, \widehat{t}\,) \le L(x, y, t)\,,$$

where $\widehat{t}$ denotes the clipped value of t at $\pm M$, that is

$$\widehat{t} := \begin{cases} -M & \text{if} \quad t < -M\,, \\ t & \text{if} \quad t \in [-M, M]\,, \\ M & \text{if} \quad t > M\,. \end{cases}$$

If L is a convex loss function, then by Steinwart and Christmann (2008, Lemma 2.23) L can be clipped at M only if $Y \subseteq [-M, M]$ and L has at least one global minimizer in $[-M, M]$. In addition, the clipping operation potentially reduces the risks, that is, $\mathcal{R}_{L,\mathrm{P}}(\widehat{f}\,) \le \mathcal{R}_{L,\mathrm{P}}(f)$. We are therefore mostly interested in bounds of risk $\mathcal{R}_{L,\mathrm{P}}(\widehat{f}\,)$ of the clipped decision function rather than the risk $\mathcal{R}_{L,\mathrm{P}}(f)$ of the unclipped decision function. This also gives us algorithmic advantages, see Steinwart et al (2011) and Chapter 5 for the case of the hinge loss and the ALS loss, respectively.

2.2 Kernels and Reproducing Kernel Hilbert Spaces

Reproducing kernels and their associated reproducing kernel Hilbert spaces (RKHS) are one of the main building blocks of SVMs, as we will see in Chapter 4 and Chapter 5. In this section, we present some basic notions of them. Let us first define kernels.

Definition 2.8 (c.f. Steinwart and Christmann (2008, Definition 4.1))**.** *Let X be a non-empty set. Then a function $k : X \times X \to \mathbb{R}$ is called a kernel on X, if there exists a Hilbert Space H and a map $\Phi : X \to H$ such that, for all $x, x' \in X$, we have*

$$k(x, x') = \langle \Phi(x), \Phi(x') \rangle_H \,, \tag{2.3}$$

where Φ is called a feature map and H a feature space of k.

In general, the feature map Φ and the feature space H are not uniquely determined, however, different feature maps and corresponding feature spaces associated to the same kernel k lead to the unique inner product $\langle \Phi(x), \Phi(x') \rangle$. For instance, we have Φ_1 and Φ_2 that map into feature spaces H_1 and H_2, respectively, associated to the same kernel k. If $\Phi_1(x) \neq \Phi_2(x)$ then consequently $H_1 \neq H_2$, and furthermore spaces H_1 and H_2 may differ in terms of their dimensions. However, we always have $\langle \Phi_1(x), \Phi_1(x') \rangle_{H_1} = \langle \Phi_2(x), \Phi_2(x') \rangle_{H_2}$. For further details in this context, we refer the reader to Schölkopf and Smola (2002, Chapter 2.2.2 and 2.2.4). Note that in case of high dimensional feature spaces, the computation of the inner product $\langle \Phi(x), \Phi(x') \rangle$ is expensive. However, for learning methods which only require the inner product of feature maps such as SVMs, the so called *kernel trick* provides an alternative way to compute inner product without knowing the feature space H and without explicitly mapping into H. In fact, the kernel trick makes it possible to compute the result of the inner product in the original space X implicitly, as we can see in the following examples of kernels. The detailed properties of these kernels can be found in Schölkopf and Smola (2002, Chapter 2.3).

Example 2.9 (Polynomial Kernel)**.** *For $m \in \mathbb{N}$, $c > 0$, and $x, x' \in \mathbb{R}^d$ for $d \geq 1$, the kernel*

$$k(x, x') := (\langle x, x' \rangle + c)^m \tag{2.4}$$

is called inhomogeneous polynomial kernel of order m. For $c = 0$, it is called homogeneous polynomial kernel. Finally, for $m = 1$ and $c = 0$, it is called linear kernel.

Example 2.10 (Exponential Kernel)**.** *For $d \in \mathbb{N}$ and $x, x' \in \mathbb{R}^d$, the kernel*

$$k(x, x') := \exp(\langle x, x' \rangle) \,, \tag{2.5}$$

is called exponential kernel.

In Definition 2.8 the feature space H is required in order to decide whether a given function k is a kernel, and this requirement sometimes becomes difficult to fulfill. In the following, we characterize kernels in terms of inequalities that helps to define kernels in a different way.

Definition 2.11 (cf. Steinwart and Christmann (2008, Definition 4.15)). *A function $k : X \times X \to \mathbb{R}$ is called **positive definite** if*

$$\sum_{i,j=1}^{n} c_i c_j k(x_i, x_j) \geq 0 \tag{2.6}$$

*holds for all $n \in \mathbb{N}$, $c_1, \ldots, c_n$, and all $x_1, \ldots, x_n \in X$. Moreover, k is said to be **strictly positive definite** if, for mutually distinct $x_1, \ldots, x_n \in X$, the equality in (2.6) only holds for $c_1 = \cdots = c_n = 0$. Finally, k is called **symmetric** if $k(x, x') = k(x', x)$ for all $x, x' \in X$.*

In the latter definition, $K := (k(x_i, x_j))_{i,j}$ for all fixed $x_1 \ldots x_n \in X$ is called the *Gram matrix*, and (2.6) is equivalent to saying that Gram matrices are positive definite. The classical and well-known result shows that the definiteness and symmetry of a function k are necessary and sufficient conditions to say that k is a kernel, see Steinwart and Christmann (2008, Theorem 4.16) for a proof. For more properties of kernel k, we refer the reader to Steinwart and Christmann (2008, Chapter 4.1).

In Definition 2.8, we note that the feature map Φ and the corresponding feature space H are not uniquely determined. One way to resolve this problem is to choose a canonical feature map of the kernel k that leads to a well-known space called the reproducing kernel Hilbert space (RKHS). This space is the smallest feature space of the kernel k in a certain sense.

Definition 2.12 (cf. Steinwart and Christmann (2008, Definition 4.18)). *Let $X \neq \emptyset$ and H be a real-valued Hilbert function space over X.*

*i) A function $k : X \times X \to \mathbb{R}$ is called **reproducing kernel** of H if we have $k(\cdot, x) \in H$ for all $x \in X$ and if the **reproducing property***

$$f(x) = \langle f, k(\cdot, x) \rangle \,,$$

holds for all $f \in H$ and all $x \in X$.

*ii) The space H is called a **reproducing kernel Hilbert space** over X if for all $x \in X$ the Dirac functional $\delta_x : H \to \mathbb{R}$ defined by*

$$\delta_x(f) := f(x) \,,$$

is continuous.

It is important to note that not every Hilbert space is RKHS but only those for which ii) holds, i.e, the Hilbert spaces in which the evaluation functionals are bounded. We further note that the reproducing kernel k is a kernel in the sense of (2.3) with feature space H and canonical feature map $\Phi : X \to H$ (see Steinwart and Christmann, 2008, Lemma 4.19). It is also known, see Steinwart and Christmann (2008, Theorem 4.20, 4.21), that reproducing kernel of a RKHS is unique, and so is the RKHS associated to the positive definite kernel.

Another way of constructing RKHSs for a continuous positive definite kernel k is to choose Mercer maps that are combinations of eigenvalues-eigenfunctions of the integral operator $T_k : L_2(X) \to L_2(X)$. To be more precise, let k be a measurable and bounded kernel on X with separable RKHS H and μ be a finite measure on X. Then the integral operator

$$(T_k f)(\cdot) := \int_X k(\cdot, x) f(x) d\mu(x) \,,$$

is compact, self-adjoint, and non-negative. Consequently, there exists an at most countable family of eigenvalues $(\lambda_i)_{i\in I} \subset (0, \infty)$ and corresponding orthonormal system (ONS) $([\tilde{e}_i]_\sim)_{i\in I} \subset L_2(\mathrm{P}_X)$ of eigenfunctions of T_k. Moreover, we have $\sum_{i\in I} \lambda_i < \infty$, and there exists a family $(e_i)_{i\in I} \in H$ with

$$[e_i]_\sim = [\tilde{e}_i]_\sim \quad \forall i \in I$$

and a measure set $N \in X$ with $P_X(N) = 0$ such that

$$k(x, x') = \sum_{i\in I} \lambda_i \cdot e_i(x) e_i(x') \,, \quad \forall x, x' \in X \setminus N \,.$$

For further details, see Steinwart and Scovel (2012, Lemma 2.1 and Corollary 3.2).

In the following, we recall the Gaussian RBF kernel and describe its associated RKHS when the input space X is a subset of $\mathbb{R}^d$. However, if X exhibits a special structure, such as text strings or DNA sequence, it is required to use a RKHS that is suitable to this structure, see e.g. Shawe-Taylor and Cristianini (2004) for a detailed overview in this context. For more technical details on the Gaussian RBF kernels and their associated RKHSs if $X \subset \mathbb{R}^d$, we refer to Steinwart and Christmann (2008, Chapter 4.4).

Definition 2.13 (cf. Steinwart and Christmann (2008, Proposition 4.10))**.** *Let $x, x' \in \mathbb{R}^d$, $d \in \mathbb{N}$. Then for all $\gamma > 0$, the $\mathbb{R}$-valued kernel*

$$k_\gamma(x, x') := \exp(-\gamma^{-2} \|x - x'\|_2^2) \,, \tag{2.7}$$

is called Gaussian RBF kernel with width γ. Here $\|\cdot\|_2$ denotes the Euclidean norm on $\mathbb{R}^d$.

Gaussian RBF kernel is translation invariant which is also referred to the stationarity of a kernel. Furthermore, a feature map $\Phi_\gamma : X \to L_2(\mathbb{R}^d)$ of the Gaussian RBF kernel k_γ for all $\gamma > 0$, see (Steinwart and Christmann, 2008, Lemma 4.45), is

$$\Phi_\gamma(x) := \left(\frac{2}{\sqrt{\pi}\gamma}\right)^{\frac{d}{2}} \exp(-2\gamma^{-2}\|x - \cdot\|_2^2) \qquad x \in X\,,$$

where $L_2(\mathbb{R}^d)$ is a feature space of k_γ. Let us denote by H_γ the RKHS of the Gaussian RBF kernel k_γ, then by Steinwart and Christmann (2008, Proposition 4.46) for any non-empty set $X \subset \mathbb{R}^d$ and $\gamma > 0$ the operator $T_\gamma : L_2(\mathbb{R}^d) \to H_\gamma(X)$

$$T_\gamma g(x) := \left(\frac{2}{\sqrt{\pi}\gamma}\right)^{\frac{d}{2}} \int_{\mathbb{R}^d} \exp(-2\gamma^{-2}\|x - y\|_2^2) g(y) dy \qquad g \in L_2(\mathbb{R}^d)\,, \quad x \in X\,,$$

is a metric surjection. In addition, by Steinwart and Christmann (2008, Theorem 4.63), the RKHS $H_\gamma(\mathbb{R}^d)$ is dense in $L_p(\mu)$ where μ is a finite measure on $\mathbb{R}^d$ and $p \in [1, \infty)$. If we restrict k_γ to $\tilde{k}_\gamma := k_{\gamma|X\times X}$ where $X \subset \mathbb{R}^d$ is a compact subset then the corresponding RKHS $\tilde{H}_\gamma$ is dense in $C(X)$ and thus $\tilde{k}_\gamma$ is a universal kernel, see also Steinwart and Christmann (2008, Lemma 4.55 and Corollary 4.58).

2.3 An Overview of the Statistical Analysis of SVMs

In this section, we give an overview of the statistical analysis of SVMs. We will also present the general oracle inequality that will serve as the basis to establish oracle inequalities and leaning rates in the case of the ALS loss in Chapter 4. For further technical details in the context of statistical analysis of SVMs, we refer to Steinwart and Christmann (2008, Chapter 4, 5, 6 & 7). Here, we recall Steinwart and Christmann (2008, Chapter 5.1) for the general SVM solution.

Definition 2.14. *Let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a loss, H be a RKHS of a measurable kernel k over X and P be a distribution on $X \times Y$. Then for $\lambda > 0$, a function $f_{\mathrm{P},\lambda} \in H$ satisfying*

$$\lambda\|f_{\mathrm{P},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda}) = \inf_{f\in H} \lambda\|f\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f)$$

*is called **general SVM solution**. Moreover, for $f_{\mathrm{P},\lambda}$ we have*

$$\lambda\|f_{\mathrm{P},\lambda}\|_H^2 \le \lambda\|f_{\mathrm{P},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda}) \le \mathcal{R}_{L,\mathrm{P}}(0).$$

It is important to know that a unique $f_{\mathrm{P},\lambda}$ exists if P is a distribution on $X \times Y$ with $\mathcal{R}_{L,\mathrm{P}}(0) < \infty$, L is a convex and locally Lipschitz continuous loss, and H is a separable RKHS of a bounded measurable kernel k over X, see Steinwart and Christmann (2008, Lemma 5.1,

Theorem 5.2 and Corollary 5.3). Since the distribution P is unknown in practice, we therefore consider the corresponding **empirical SVM solution**, see e.g. Steinwart and Christmann (2008, Theorem 5.5).

Theorem 2.15 (Representer Theorem)**.** *Let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a convex loss, H be a RKHS over X and $D := ((x_1, y_n), \ldots, (x_n, y_n)) \in (X \times Y)^n$ be a data set. Then for all $\lambda > 0$, there exits a unique $f_{D,\lambda} \in H$ such that*

$$\lambda \|f_{D,\lambda}\|_H^2 + \mathcal{R}_{L,D}(f_{D,\lambda}) = \inf_{f \in H} \lambda \|f\|_H^2 + \mathcal{R}_{L,D}(f)\,. \tag{2.8}$$

In addition, there exist $\alpha_1, \ldots \alpha_n \in \mathbb{R}$ such that

$$f_{D,\lambda}(\cdot) = \sum_{i=1}^{n} \alpha_i k(x_i, \cdot)\,. \tag{2.9}$$

If L is a convex loss and H is a separable RKHS, then the decision function $f_{D,\lambda}$ for all $\lambda > 0$ and the corresponding learning method producing $f_{D,\lambda}$ are measurable, see Steinwart and Christmann (2008, Lemma 6.23). Additionally, if L is a continuous loss that is differentiable, then the maps $D \mapsto f_{D,\lambda}$ mapping $(X \times Y)^n$ to H are continuous, see Steinwart and Christmann (2008, Lemma 5.13). From (2.9) we further see that the decision function $f_{D,\lambda}$ elucidates the importance of kernels. In other words, by transferring the solution $f_{D,\lambda}$ into a kernel representation, often called dual representation with dual variables $\alpha \in \mathbb{R}^n$, one can reduce the computational efforts in applications. We will elaborate the general idea of dual formulation of an optimization problem in Section 2.4 and the computation of dual variables in the context of ALS loss in Chapter 5. To this end, we return to the idea of the general SVM solution $f_{P,\lambda}$. If $L : X \times Y \times \mathbb{R} \to [0, \infty)$ is a convex, P-integrable Nemitski loss of order $p \in [1, \infty)$, then by Steinwart and Christmann (2008, Chapter 5.2) and Steinwart and Christmann (2008, Theorem 5.8), the kernel representation of $f_{P,\lambda}$ is

$$f_{P,\lambda}(\cdot) = -\frac{1}{2\lambda} \int_{X \times Y} h(x, y) k(x, \cdot) dP(x, y) = -\frac{1}{2\lambda} \mathbb{E}_P h \Phi\,, \tag{2.10}$$

where $h(x, y) \in \partial L(x, y, f_{P,\lambda}(x))$, $(x, y) \in X \times Y$ and $\partial L(\cdot)$ denotes the subdifferential of L w.r.t. the third argument, see (Steinwart and Christmann, 2008, Lemma A.6.15) for further details. Similar to (2.10), we now reformulate the kernel representation of the empirical SVM solution $f_{D,\lambda}$ (2.9), that is

$$f_{D,\lambda}(\cdot) = -\frac{1}{2n\lambda} \sum_{i=1}^{n} h(x_i, y_i) k(x_i, \cdot) = -\frac{1}{2\lambda} \mathbb{E}_D h \Phi\,, \tag{2.11}$$

where $h(x,y) \in \partial L(x,y,f_{\mathrm{D},\lambda}(x))$ for all $(x,y) \in X \times Y$. From (2.11) we see that the possible dual coefficients $\alpha_i, i = 1,\ldots,n$ are determined by

$$\alpha_i := \frac{h(x_i,y_i)}{2n\lambda}\,, \qquad i = 1,\ldots,n\,.$$

Both the decision function $f_{\mathrm{D},\lambda}$ produced by the SVM in Theorem 2.15 and the associated risk $\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda})$ are random variables because data D in general comprises i.i.d. observations from some unknown distribution P. Therefore, for an $f_{\mathrm{D},\lambda}$, one is usually interested to determine *learning ability* of an SVM. In other words, one wants to know that with what probability, the risk $\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda})$ is close to the Bayes' risk $\mathcal{R}^*_{L,\mathrm{P}}$. One way to address this question is to establish the *L-risk consistency* for P, see Steinwart and Christmann (2008, Definition 6.4), that is

$$\lim_{n\to\infty} P^n(D \in (X \times Y)^n : \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda}) \le \mathcal{R}^*_{L,\mathrm{P}} + \varepsilon) = 1\,, \tag{2.12}$$

for all $\varepsilon > 0$. Moreover, (2.12) leads to universal L-risk consistency, if it is L-risk consistent for all distributions P on $X \times Y$. For universal consistency of learning methods for binary classification and least squares regression, we refer to Devroye et al (1996) and Györfi et al (2002), respectively. Clearly, the consistency definition (2.12) does not specify the speed of convergence of the learning method. Therefore, a better approach is to establish *learning rates*, see, e.g. Steinwart and Christmann (2008, Lemma 6.5). To be more precise, for a fixed sequence $(\varepsilon_n) \subset (0,1]$ that converges to 0, we say that the learning method learns with rate (ε_n), if there exists a family $(c_\varrho)_{\varrho\in(0,1]}$ such that for all $n \ge 1$ and all $\varrho \in (0,1]$, we have

$$P^n(D \in (X \times Y) : \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda}) \le \mathcal{R}^*_{L,\mathrm{P}} + c_{\mathrm{P}}\, c_\varrho\, \varepsilon_n) \ge 1 - \varrho\,. \tag{2.13}$$

Note that learning rate (2.13) includes a constant c_{P} that depends on the unknown data generating distribution P , and by no-free-lunch theorem (see, e.g. Devroye et al, 1996, Theorem 7.2) there exists no learning method that enjoys uniform learning rates for all distributions P. One way to cope with this issue is to make a *priori* assumptions on the distribution P, that is, by establishing learning rates under different assumptions on P, one can explore the distributions for which learning method learns well.

Recall that the statistical analysis of both empirical risk minimization (ERM), see Steinwart and Christmann (2008, Chapter 6.3) for further details, and SVMs relies on bounds of the probabilities

$$P^n(D \in (X \times Y)^n : |\mathcal{R}_{L,\mathrm{D}}(f_{\mathrm{D},\lambda}) - \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda})| > \varepsilon)\,.$$

In order to establish these bounds, well-known *concentration inequalities* such as Markov's inequality, Hoeffding's inequality, Berstein's inequality and Talagrand's inequality are given in (Steinwart and Christmann, 2008, Chapter 6.2 and Appendix A.9). These lead to oracle inequalities for SVMs where each relates the risk of an empirical SVM solution to the corresponding infinite-sample SVM. For more details on oracle inequalities for SVMs, we refer to Steinwart and Christmann (2008, Chapter 6.4 & Chapter 7.4). In this section, we will only recall the general oracle inequality for SVMs established in (Steinwart and Christmann, 2008, Theorem 7.23). In order to fully understand this oracle inequality, we first recall notions of supremum bound and variance bound of a loss function L.

Definition 2.16. *Let $L : X \times Y \times \mathbb{R} \to \mathbb{R}$ be a loss that can be clipped at some $M > 0$ and P be a distribution on $X \times Y$ such that the Bayes decision function $f^*_{L,\mathrm{P}} : X \to [-M, M]$ exists. Then we say that L satisfies a supremum bound*

$$\|L \circ f - L \circ f^*_{L,\mathrm{P}}\|_\infty \leq B\,, \tag{2.14}$$

if there exists a constant $B > 0$. In addition, for all $(x, y) \in X \times Y$ and $f : X \to [-M, M]$, L satisfies a variance bound

$$\mathbb{E}_\mathrm{P}(L \circ f - L \circ f^*_{L,\mathrm{P}})^2 \leq V(\mathbb{E}_\mathrm{P}(L \circ f - L \circ f^*_{L,\mathrm{P}}))^\vartheta\,, \tag{2.15}$$

if there exists a $\vartheta \in (0, 1)$ such that $V \geq B^{2-\vartheta}$.

For many loss functions, establishing a variance bound (2.15) is a non-trivial task. We refer the reader to Steinwart and Christmann (2008, Theorem 8.24), Steinwart and Christmann (2008, Example 7.3) and Steinwart and Christmann (2011, Theorem 2.8) for variance bounds of hinge loss, least squares loss and pinball loss, respectively. Moreover, the variance bounds for the ALS loss are established in Chapter 4.

We now introduce the concept of *covering numbers*, see e.g. Steinwart and Christmann (2008, Definition 6.19) which is used to control the capacity of the underlying RKHS H.

Definition 2.17. *For a metric space $(\mathcal{F}, d)$ and $\varepsilon > 0$, a subset $S \subset \mathcal{F}$ is called an ε-net of $\mathcal{F}$ if for all $f \in \mathcal{F}$ there exists an $s \in S$ with $d(s, f) \leq \varepsilon$. Moreover, the ε-covering number of $\mathcal{F}$ is defined by*

$$\mathcal{N}(\mathcal{F}, d, \varepsilon) := \inf\left\{n \geq 1 : \exists\, s_1, \ldots s_n \in \mathcal{F} \text{ such that } \mathcal{F} \subset \bigcup_{i=1}^n B_d(s_i, \varepsilon)\right\},$$

where $\inf \emptyset := \infty$ and $B_d(s, \varepsilon) := \{f \in \mathcal{F} : d(f, s) \leq \varepsilon\}$ denotes the closed ball with center $s \in \mathcal{F}$ and radius ε.

The covering number $\mathcal{N}(\mathcal{F}, d, \varepsilon)$ is in fact the size of the smallest possible ε-net that is needed to approximate the set $\mathcal{F}$ with accuracy ε. Another way to control the capacity of RKHSs is the entropy numbers, see e.g.Steinwart and Christmann (2008, Definition 6.20), which is the dual of the covering numbers..

Definition 2.18 (Entropy number)**.** *Let $(\mathcal{F}, d)$ be a metric space and $n \geq 1$ be an integer. Then the n-th (dyadic) entropy number of $(\mathcal{F}, d)$ is defined by*

$$e_n(\mathcal{F}, d) := \inf \left\{ \varepsilon > 0 : \exists\, s_1, \dots s_n \in \mathcal{F} \text{ such that } \mathcal{F} \subset \bigcup_{i=1}^{2^{n-1}} B_d(s_i, \varepsilon) \right\}.$$

Moreover, let $T : E \to F$ be a bounded, linear operator between the normed spaces E and F, then $e_i(T) := e_i(TB_E, \|\cdot\|_F)$.

Note that the bounds on entropy numbers imply equivalent bounds on covering numbers and vice verse, as shown in Steinwart and Christmann (2008, Lemma 6.21) and Steinwart and Christmann (2008, Exercise 6.8). In the following lemma, we present one directional relation.

Lemma 2.19. *Let $(\mathcal{F}, d)$ be a metric space, $c > 0$ and $p > 0$ be constants such that*

$$\ln \mathcal{N}(\mathcal{F}, d, \varepsilon) < \left(\frac{c}{\varepsilon}\right)^p,$$

for all $\varepsilon > 0$. Then $e_n(\mathcal{F}, d) \leq 3^{\frac{1}{p}}\, c\, n^{-\frac{1}{p}}$ for all $n \geq 1$.

Let us now present a general oracle inequality for SVMs that is given in Steinwart and Christmann (2008, Theorem 7.23). This will provide the basis to establish oracle inequalities and corresponding learning rates in the case of the ALS loss in Chapter 4.

Theorem 2.20 (Oracle inequality for SVMs)**.** *Let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a locally Lipschitz continuous loss that can be clipped at $M > 0$ and satisfies the supremum bound* (2.14) *for a $B > 0$. Moreover, let H be a separable RKHS of a measurable kernel k over X and P be a distribution on $X \times Y$ such that the variance bound* (2.15) *is satisfied for constants $\vartheta \in [0, 1]$, $V \geq B^{2-\vartheta}$, and all $f \in H$. Assume that for fixed $n \geq 1$, there exist constants $p \in (0, 1)$ and $a \geq B$ such that*

$$\mathbb{E}_{D \sim P_X^n} e_n(\mathrm{id} : H \to L_2(D_X)) \leq a n^{-\frac{1}{2p}}, \qquad i \geq 1. \tag{2.16}$$

Finally, fix an $f_0 \in H$ and a constant $B_0 \geq B$ such that $|L \circ f_0|_\infty \leq B_0$. Then, for all fixed $\varrho > 0$ and $\lambda > 0$, the SVM using H and L satisfies

$$\lambda \|f_{\mathrm{D},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{D,\lambda}) - \mathcal{R}^*_{L,\mathrm{P}} \leq 9(\lambda \|f_0\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f_0) - \mathcal{R}^*_{L,\mathrm{P}})$$

$$+ K\left(\frac{a^{2p}}{\lambda^p n}\right)^{\frac{1}{2-p-\vartheta+\vartheta p}} + 3\left(\frac{72V\varrho}{n}\right)^{\frac{1}{2-\vartheta}} + \frac{15B_0\varrho}{n}\,, \qquad (2.17)$$

with probability P^n *not less than* $1-3e^{-\varrho}$*, where* $K \geq 1$ *is a constant only depending on* p*,* M*,* ϑ *and* V*.*

If k is a Gaussian kernel, the constant K in (2.17) depends on p in an unknown manner, see (Steinwart and Christmann, 2008, Theorem 7.16). In Chapter 4 we will show an explicit bound for K considering $p \in (0, \frac{1}{2}]$. The right hand side of the oracle inequality (2.17) consists of two parts, namely the approximation error and the estimation error. If the distribution P is such that $\mathcal{R}^*_{L,\mathrm{P}} < \infty$ holds, then the approximation error function $\mathcal{A} : [0,\infty) \to [0,\infty)$, see Steinwart and Christmann (2008, Definition 5.14), is defined by

$$\mathcal{A}(\lambda) := \inf_{f\in H} \lambda\|f\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}} < \infty\,, \qquad \lambda \geq 0\,, \qquad (2.18)$$

The approximation error function $\mathcal{A}(\lambda)$ is increasing, concave and continuous (see, Steinwart and Christmann, 2008, Lemma 5.15)). By Steinwart and Christmann (2008, Corollary 5.18), there exists a constant $c > 0$ such that the approximation error function can be bounded by $\mathcal{A}(\lambda) \leq c\,\lambda$ for all $\lambda > 0$ if and only if $f^*_{\mathrm{P},\lambda} \in H$.

2.4 Introduction to Convex Optimization

This section contains an overview of some of the basic tools that are required to solve the optimization problem (1.7). In particular, we will deal with constrained convex optimization problems. In addition, we will give a brief overview on optimization algorithms to deal with such problems. The contents of this section mainly follow Schölkopf and Smola (2002, Chapter 6), Cristianini and Shawe-Taylor (2000, Chapter 5), Steinwart and Christmann (2008, Chapter 11) and Boyd and Vandenberghe (2004). Let us begin by the definition of the primal optimization problem, see e.g Cristianini and Shawe-Taylor (2000, Definition 5.1), Schölkopf and Smola (2002, Chapter 6.3) and Abe (2005, Chapter 5.5.1).

Definition 2.21. *Let* f*,* $g_i, i = 1,\dots k$ *and* $h_j, j = 1,\dots \ell$ *be functions defined on a domain* $\Omega \subseteq \mathbb{R}^n$*. Then a* ***primal optimization problem*** *(*$\mathcal{P}$*) is of the form:*

$$\min_{w\in\Omega} \quad f(\mathbf{w}) \qquad (2.19)$$

$$\text{subject to} \quad g_i(\mathbf{w}) \leq 0\,, \qquad i = 1,\dots,k\,, \qquad (2.20)$$

$$h_j(\mathbf{w}) = 0\,, \qquad j = 1,\dots,\ell\,, \qquad (2.21)$$

The function $f(\mathbf{w})$ in (2.19) is the objective function, and (2.20) and (2.21) are inequality and equality constraints, respectively. In general, there exists the region $\mathcal{M}$ of the domain $\Omega \subseteq \mathbb{R}^n$ called *feasible region*

$$\mathcal{M} := \{\mathbf{w} \in \Omega : \mathbf{g}(\mathbf{w}) \leq \mathbf{0}, \mathbf{h}(\mathbf{w}) = \mathbf{0}\}, \tag{2.22}$$

that contain the solution, either local or global, of the optimization problem $\mathcal{P}$ if $\mathcal{M} \neq \emptyset$. In other words, the solution $\mathbf{w}^* \in \mathcal{M}$ is called the *global minimum* if there exists no other $\mathbf{w} \in \mathcal{M}$ for which $f(\mathbf{w}) < f(\mathbf{w}^*)$ holds and as a result $f(\mathbf{w}^*)$ is called the optimal value of $\mathcal{P}$. On the other hand, $\mathbf{w}^* \in \mathcal{M}$ is called a *local minimum* if there is an $\varepsilon > 0$ with $f(\mathbf{w}) \geq f(\mathbf{w}^*)$ for all for all $\mathbf{w} \in \mathcal{M}$ with $\|\mathbf{w} - \mathbf{w}^*\| < \varepsilon$. Note that if all constraints are linear and the objective function is quadratic, then the optimization problem is called *quadratic programming*. However, it is called *linear programming* if the objective function is linear too. The optimization problem is said to be *convex* if the objective function and all the constraints are convex. In the following, we will always consider the quadratic optimization problem which is convex too and refer to Steinwart and Christmann (2008, Appendix A.6) for basic properties of convex functions. The main reasons for considering aforementioned problem are, first, it leads to a unique global solution, see Schölkopf and Smola (2002, Theorem 6.11) and Steinwart and Christmann (2008, A.6.9)), and secondly, there are many efficient algorithms available to solve convex quadratic programs, that we will discuss briefly later in this section.

To this end, we define the Lagrangian function that is a key ingredient to find the solution of an optimization problem.

Definition 2.22. *Consider the optimization problem $\mathcal{P}$ where $f, g_i, h_j : \mathbb{R}^n \to \mathbb{R}$ for $i = 1, \ldots, k$ and $j = 1, \ldots, \ell$. Then the* ***Lagrangian function*** $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^k \times \mathbb{R}^\ell \to \mathbb{R}$ *is defined by*

$$\mathcal{L}(\mathbf{w}, \alpha, \beta) := f(\mathbf{w}) + \sum_{i=1}^{n} \alpha_i g_i(\mathbf{w}) + \sum_{j=1}^{m} \beta_j h_j(\mathbf{w}), \tag{2.23}$$

where $\alpha_i \in [0, \infty)$ for $i = 1, \ldots, n$ and $\beta_j \in \mathbb{R}$ for $j = 1, \ldots, m$ are called ***Lagrange multipliers*** *or* ***dual variables*** *associated with problem $\mathcal{P}$.*

Based on the the Lagrangian function $\mathcal{L}$, we now transform the primal optimization problem $\mathcal{P}$ into the Lagrangian dual optimization problem.

Definition 2.23. *Let $\mathcal{P}$ be a convex problem and $\mathcal{L}$ be the Lagrangian function. Moreover, define the function $\mathcal{D} : \mathbb{R}^n \to \mathbb{R}$ by*

$$\mathcal{D}(\alpha, \beta) := \inf_{\mathbf{w} \in \mathbb{R}^n} \mathcal{L}(\mathbf{w}, \alpha, \beta). \tag{2.24}$$

Then the ***dual optimization problem*** *is*

$$\begin{aligned} \arg\max_{\alpha,\beta} \quad & \mathcal{D}(\alpha,\beta)\,, \\ \text{subject to} \quad & \alpha \geq 0\,. \end{aligned} \tag{2.25}$$

Note that the dual function $\mathcal{D}$ is concave even when the primal $\mathcal{P}$ is not convex, because the dual function $\mathcal{D}$ is the point-wise infimum of a family of *affine* functions of (α, β). Furthermore, for each pair (α, β) with $\alpha \geq 0$, the Lagrange dual function $\mathcal{D}(\alpha, \beta)$ gives a lower bound on the optimal value $f(\mathbf{w}^*)$ of the optimization problem $\mathcal{P}$, where the optimal value of the Lagrange dual problem $\mathcal{D}(\alpha^*, \beta^*)$ is the best lower bound on $f(\mathbf{w}^*)$. This relationship between the solution of the primal and the dual problem can be explained by the notion of weak duality theorem, see e.g. Cristianini and Shawe-Taylor (2000, Theorem 5.15).

Theorem 2.24 (cf. Cristianini and Shawe-Taylor (2000, Theorem 5.15 and Corollary 5.16))**.** *Let* $\mathbf{w} \in \mathbb{R}^n$ *be a feasible solution of a primal problem* $\mathcal{P}$ *and* (α, β) *be a feasible solution of a dual problem* $\mathcal{D}$*. Then*

$$f(\mathbf{w}) \geq \mathcal{D}(\alpha, \beta)\,.$$

Moreover, the value of the dual $\mathcal{D}$ *is upper bounded by the value of the primal* $\mathcal{P}$

$$\sup\{\mathcal{D}(\alpha, \beta) : \alpha \geq \mathbf{0}\} \leq \inf\{f(\mathbf{w}) : \mathbf{g}(\mathbf{w}) \leq \mathbf{0}, \mathbf{h}(\mathbf{w}) = \mathbf{0}\}\,. \tag{2.26}$$

Clearly, we see from (2.26) that the difference between values of the primal and the dual problems may exists and this difference is called the *duality gap*. Note that this duality gap, in fact, may serve as a stopping criterion of the algorithm used to solve the optimization problem. We refer the reader to Chapter 5 to see the technical details for establishing the duality gap in the case of the ALS loss. To this end, we present the notion of strong duality, which guarantees that the dual and the primal problems reach the same solution for an optimization problem.

Theorem 2.25 (cf. Cristianini and Shawe-Taylor (2000, Theorem 5.20))**.** *Given an optimization problem* $\mathcal{P}$ *with convex domain* $\Omega \subset \mathbb{R}^n$*, and affine functions* g_i *and* h_j *in* (2.20) *and* (2.21) *respectively, that is*

$$\mathbf{h}(\mathbf{w}) = \mathbf{A}\mathbf{w} - \mathbf{b}\,,$$

holds for some matrix $\mathbf{A}$ *and vector* $\mathbf{b}$*, the duality gap is zero.*

Note that if $\mathbf{w}^*$ and (α^*, β^*) are the primal and the dual optimal solution such that the strong duality holds, then $(\mathbf{w}^*, \alpha^*, \beta^*)$ form a *saddle-point* for the Lagrangian function $\mathcal{L}$, and

vise versa. To be more precise, the vector $(\mathbf{w}^*, \alpha^*, \beta^*)$ is called a saddle point of $\mathcal{L}(\mathbf{w}, \alpha, \beta)$ with respect to maximizing in α and β, and minimizing in $\mathbf{w}$ if for all $\mathbf{w} \in \mathbb{R}^n$, $\alpha \in [0, \infty)^k$ and $\beta \in \mathbb{R}^\ell$, we have

$$\mathcal{L}(\mathbf{w}^*, \alpha, \beta) \leq \mathcal{L}(\mathbf{w}^*, \alpha^*, \beta^*) \leq \mathcal{L}(\mathbf{w}, \alpha^*, \beta^*) \tag{2.27}$$

In the following, we give the Kuhn-Tucker conditions for an optimum solution to an optimization problem.

Theorem 2.26 (cf. Cristianini and Shawe-Taylor (2000, Theorem 5.21))**.** *Let $\mathcal{P}$ be a convex optimization problem with $\Omega \subset \mathbb{R}^n$. Moreover, assume that f is convex function and g_i, h_j are affine. Then the necessary and sufficient conditions for a normal point $\mathbf{w}^*$ to be an optimum are the existence of α^* and β^* such that*

$$\frac{\partial \mathcal{L}(\mathbf{w}^*, \alpha^*, \beta^*)}{\partial \mathbf{w}} = 0\,, \tag{2.28}$$

$$\frac{\partial \mathcal{L}(\mathbf{w}^*, \alpha^*, \beta^*)}{\partial \beta} = 0\,, \tag{2.29}$$

$$\alpha_i^* g_i(\mathbf{w}^*) = 0\,, \quad i = 1, \ldots, k\,, \tag{2.30}$$

$$g_i(\mathbf{w}^*) \leq 0\,, \quad i = 1, \ldots, k\,, \tag{2.31}$$

$$\alpha_i^* \geq 0\,, \quad i = 1, \ldots, k\,. \tag{2.32}$$

Here, the third condition (2.30) is called Karush–Kuhn–Tucker (KKT) complementary condition which implies that for an active constraint, the dual variable $\alpha_i^* > 0$ holds, and for an inactive constraint we have $\alpha_i^* = 0$. However, perturbing inactive constraints have no influence on the solution of optimization problem.

The standard approach of solving an SVM problem basically solves a convex optimization problem of the form shown in Definition 2.21 using sample information, which we further transform into dual form like Definition 2.23, see e.g. Cristianini and Shawe-Taylor (2000, Chapter 6) and Steinwart and Christmann (2008, Chapter 11.1). There are several standard numerical methods to solve the convex optimization problem like descent methods, interior-point methods, decomposition methods, etc. We refer the interested reader to Cristianini and Shawe-Taylor (2000, Chapter 7), Schölkopf and Smola (2002, Chapter 10), Boyd and Vandenberghe (2004) and Steinwart and Christmann (2008, Chapter 11.2) for detailed survey on different convex optimization methods. In the following, we mainly focus on the *decomposition methods* (also called *chunking* or *subset selection method*) which is used in case of large sample size. The main idea of this method is to break the optimization problem into smaller subproblems and

then solve each subproblem in an iterative manner. In other words, a subset of dual variable, generally called *working set* or *active set*, is updated in each iteration in contrast to other convex optimization methods like descent methods or interior-point methods, where the whole vector of dual variable is updated in each step. In additoin, this method does not require to store the kernel matrix into the RAM of computer. However, updating few components at each iteration can increase overall computation time when the sample size is very large due to the requirement of many iterations for convergence. Hence, it is important to choose a working set in a very smart way, e.g., choosing the points that contribute most to the duality gap or those that most violate the KKT conditions, such that the optimization of the sub-problem leads to an improvement in the overall objective function.

The extreme case of the decomposition method proposed by Platt (1999) is sequential minimal optimization (SMO) method where the optimization step is done for working set of size two. In other words, at each step, SMO optimizes two chosen points of dual vector keeping fixed all others, and then updates the whole dual vector accordingly. Interestingly, the optimization problem in case of SVM for two points can often be calculated analytically. This in return, despite needing more iterations to converge, saves a lot of computation time by eliminating the need of calling an iterative convex program optimizer at each iteration step as required by general decomposition methods. Note that, the standard SVM optimization problem includes a bias term and leads to a condition which is only fulfilled if two dual points are updated at each iteration. On the other hand, SVMs without bias term eliminate this condition and thus we can establish an SMO algorithm for one working set. In order to see technical details of establishing SMO-type algorithms for one and two working set in the case of the ALS loss, we refer to Chapter 5 and also refer to Steinwart et al (2011). for the case of hinge loss.

Chapter 3

Asymmetric Least Squares Loss: Self-Calibration and Variance Bounds

The goal of this chapter is to characterize the asymmetric least square (ALS) loss function. To be more precise, we investigate some properties for the case of the ALS loss such as Lipschitz continuity, a self-calibration inequality, a supremum bound, and a variance bound. To the best of our knowledge, these properties have not been investigated in the literature. With the help of these properties, oracle inequalities and learning rates will be established for SVM-type learning algorithm for expectile regression in Chapter 4.

3.1 Loss Functions for Quantiles and Expectiles

There exists a class of loss functions that are consistent to quantiles. A general form of such loss functions can be found in Gneiting (2011, Theorem 9), see also Steinwart et al (2014, Equation (21)). However, the only loss function that has been used in the literature in order to estimate (conditional) quantiles is the asymmetric least absolute deviation (ALAD) loss function, see, e.g Koenker and Bassett Jr (1978). For all $t \in \mathbb{R}$ and any $\alpha \in (0,1)$, the ALAD loss is defined by

$$L_\alpha(y,t) = \begin{cases} (1-\alpha)|y-t|\,, & \text{if} \quad y < t\,, \\ \alpha|y-t|\,, & \text{if} \quad y \geqslant t\,. \end{cases} \tag{3.1}$$

Note that L_α- loss is convex and continuous but its first derivative with respect to second argument does not exist at $t = 0$. Consequently, the loss (3.1) makes the the analytic solution of optimization problem very challenging.

Analogous to quantiles, there also exits a class of loss functions for expectiles. We refer to

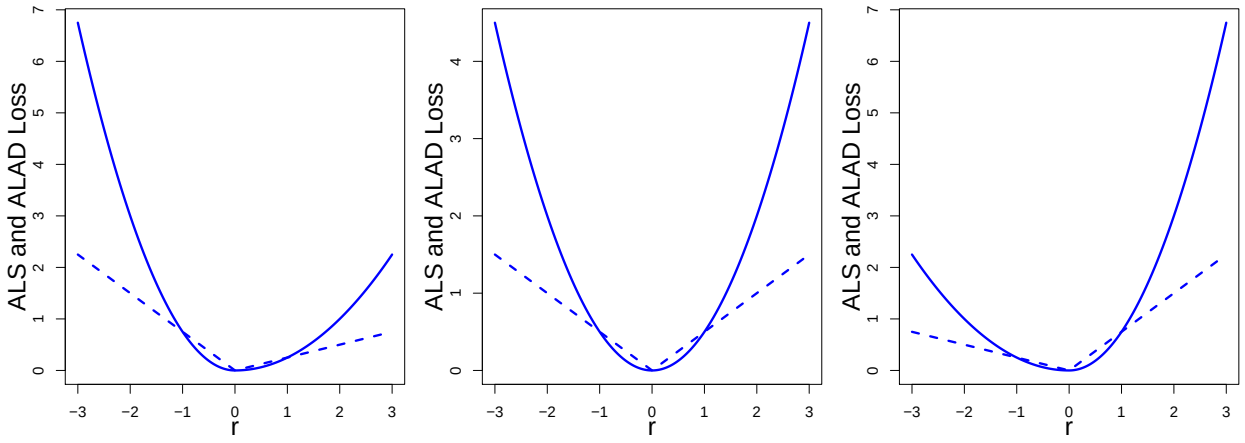

Figure 3.1: The ALS loss (solid lines) and the ALAD loss (dotted lines) for $\tau\,(\text{and}\,\alpha) = 0.25$ (left), $\tau\,(\text{and}\,\alpha) = 0.50$ (middle) and $\tau\,(\text{and}\,\alpha) = 0.75$ (right) considering $r := (y - t) \in [-3, 3]$.

Gneiting (2011, Theorem 10) and Steinwart et al (2014, Equation (26)) for the general form of such loss functions. Here it is also interesting to note that only the ALS loss proposed by Newey and Powell (1987) has been used in the literature so far in order to estimate (conditional) expectiles. Recall Newey and Powell (1987) that for any $\tau \in (0, 1)$ and all $t \in \mathbb{R}$, the ALS loss is defined by

$$L_\tau(y, t) = \begin{cases} (1-\tau)(y-t)^2\,, & \text{if} \quad y < t\,, \\ \tau(y-t)^2\,, & \text{if} \quad y \geqslant t\,. \end{cases} \tag{3.2}$$

The L_τ-loss is also convex and continuous. In addition, it is is differentiable at all t. Therefore the optimization problem based on (3.2) can be solved with the help of gradient-based methods. The illustration of both the L_α-loss and the L_τ-loss for $\tau = \alpha = 0.25, 0.5, 0.75$ is given in Figure 3.1.

It is interesting to note that the L_α-loss and the L_τ-loss behave differently. Therefore, the resultant expectiles and quantiles for $\tau = \alpha \in (0, 1)$ based on their respective losses, in general, do not coincide to each others. We further refer to Jones (1994) where it has been shown that expectiles of a distribution, say Q, coincide with quantiles of some other distribution that is related to Q. This fact is illustrated in Figure 3.2 for some standard distributions. In addition, Figure 3.3 illustrates the expectiles and quantiles of student t-distribution, where we notice that for student t-distribution with 2 degree of freedom, expectiles and quantiles coincide. The rescaled version of this distribution is derived by Koenker (1992) and named it the *Koenker distribution*. However, note that, the variance of both of the aforementioned distributions is not finite, which is considered a crucial assumption for statistical analysis for SVM-tpye leaning algorithm for expectile regression, see Chapter 4 for further details.

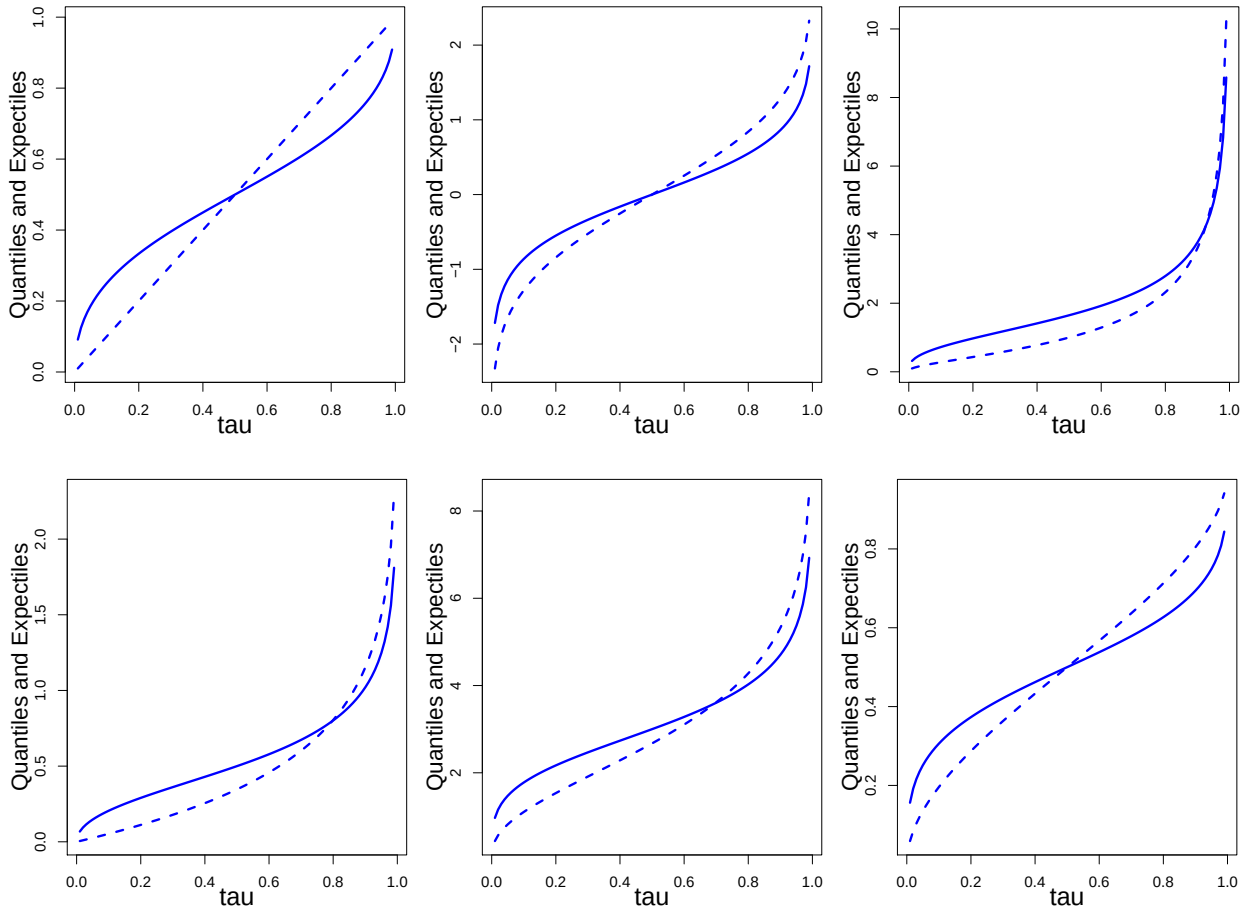

Figure 3.2: Expectiles (solid lines) and quantiles (dotted lines) for $\tau = \alpha \in (0,1)$ for uniform distribution $U(0,1)$ (top-left), normal distribution $N(0,1)$ (top-middle), lognormal distribution $\log N(0,1)$ (top-right), exponential distribution $\mathsf{exp}(\lambda = 1)$ (bottom-left), gamma distribution $\mathsf{Gamma}(\alpha = 3, \beta = 1)$ (bottom-middle) and beta distribution $\mathsf{Beta}(\alpha = 2, \beta = 2)$ (bottom-right).

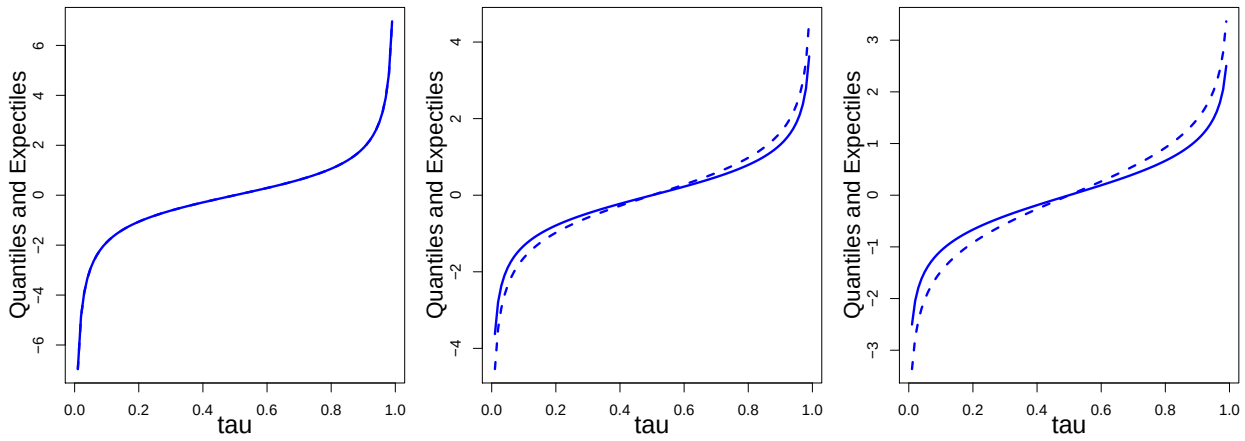

Figure 3.3: Expectiles (solid lines) and quantiles (dotted lines) for $\tau = \alpha \in (0,1)$ for student t distribution with df $= 2$ (left), df $= 3$ (middle) and df $= 5$(right).

In Figure 3.1 we also see that the L_τ-loss is more sensitive to the extreme values. It was also observed by Koenker (2005) that expectiles have more global dependence on the shape of the distribution compared to quantiles. This means that altering the one tail of distribution does not effect the quantiles of the other tail, but it does effect all the expectiles of the distribution. For more properties of expectiles, we refer the interested reader to Newey and Powell (1987), Jones (1994) and Abdous and Remillard (1995). In the following, we present some properties of L_τ. To the best of our knowledge, most of them have not been investigated in the literature.

3.2 Properties of the Asymmetric Least Squares Loss

Throughout this section, we assume that X is an arbitrary, non-empty set equipped with a σ-algebra, and $Y \subset \mathbb{R}$, if it is not stated otherwise, is a closed non-empty set. In addition, we assume that P is the probability distribution on $X \times Y$ satisfies $|\mathrm{P}|_2 := \left(\int_{X\times Y} y^2 \, d\mathrm{P}(x,y) \right)^{1/2} < \infty$, $\mathrm{P}(\,\cdot\,|x)$ is a regular conditional distribution on Y given $x \in X$ and Q is some distribution on Y. Furthermore, we assume that $L_\tau : Y \times \mathbb{R} \to [0,\infty)$ is the ALS loss defined by (3.2) and $f : X \to \mathbb{R}$ is a measurable function.

3.2.1 Convexity

It is trivial to show that L_τ for each $\tau \in (0,1)$ is convex in t, see also Figure 3.1 for illustration. This convexity further ensures that the optimization problem (1.7) is efficiently solvable. In addition, by (Steinwart and Christmann, 2008, Lemma 2.13) convexity of L_τ implies convexity of the corresponding risk. By Definition 2.3, given a predictor $f : X \to \mathbb{R}$ the L_τ-risk for each $\tau \in (0,1)$ is defined by

$$\mathcal{R}_{L_\tau,\mathrm{P}}(f) := \int_{X\times Y} L_\tau(y, f(x)) \mathrm{dP}(x,y) = \int_X \int_Y L_\tau(y, f(x)) \mathrm{dP}(y|x) \mathrm{dP}_X(x) \tag{3.3}$$

and the minimal L_τ-risk is defined by

$$\mathcal{R}^*_{L_\tau,\mathrm{P}} := \inf\{\mathcal{R}_{L_\tau,\mathrm{P}}(f) | f : X \to \mathbb{R} \text{ is measurable}\}\,. \tag{3.4}$$

The integral over $X \times Y$ always exists because $(x,y) \mapsto L_\tau(y, f(x))$ is measurable and non-negative. Moreover, one can easily show that $\mathcal{R}_{L_\tau,\mathrm{P}}(0) < \infty$ since $|\mathrm{P}|_2 < \infty$.

3.2.2 Clipping

Let us assume that $Y \subseteq [-M, M]$ for some $M > 0$, then convexity of L_τ ensures that $L_\tau(y,\cdot)$ leads to a global minimizer in Y for all $y \in Y \subseteq [-M, M]$, see Steinwart and Christmann (2008,

Lemma 2.23). This further implies that L_τ can be clipped at M in the sense of Definition 2.7. Note that the clipping assumption has already been utilized while establishing learning rates for SVMs, see for instance Chen et al (2004); Steinwart et al (2011) for hinge loss, and Christmann and Steinwart (2007) and Steinwart and Christmann (2011) for pinball loss. If we denote by $\widehat{f} : X \to [-M, M]$ the clipped decision function, then we have $\mathcal{R}_{L_\tau,\mathrm{P}}(\widehat{f}) \leq \mathcal{R}_{L_\tau,\mathrm{P}}(f)$ for every $f : X \to \mathbb{R}$. In other words, the clipping operation potentially reduces the risks. In Chapter 4, we will therefore bound the risk $\mathcal{R}_{L_\tau,\mathrm{P}}(\widehat{f}_\mathrm{D})$ of the clipped decision function rather than the risk $\mathcal{R}_{L_\tau,\mathrm{P}}(f_\mathrm{D})$, where f_D is the decision function obtained by solving optimization problem of the form (1.7) . In practice, see Chapter 5, the *training* algorithm for (1.7) remains unchanged and the *evaluation* of the resulting decision function requires only a slight change. For further details on algorithmic advantages of clipping for SVMs using the hinge loss and the ALS loss, we refer the reader to Steinwart et al (2011) and Chapter 5, respectively.

3.2.3 Local Lipschitz Continuity

Recall Definition 2.4 that a loss function is called locally Lipschitz continuous if for all $a \geq 0$ there exists a constant c_a such that

$$\sup_{y \in Y} |L(y,t) - L(y,t')| \leq c_a \, |t - t'| \,, \qquad t, t' \in [-a, a] \,.$$

In the following we consider $a := M$ and denote for a given $M > 0$ the smallest such constant c_a by $|L|_{1,M}$, and show that the ALS loss is locally Lipschitz continuous.

Lemma 3.1. *Let $Y \subseteq [-M, M]$ with $M > 0$ and $t \in Y$, then the loss function $L_\tau : Y \times [-M, M] \to [0, \infty)$ is locally Lipschitz continuous with Lipschitz constant*

$$|L_\tau|_{1,M} = C_\tau \, 4M \,,$$

where $C_\tau := \max\{\tau, 1 - \tau\}$.

Proof of Lemma 3.1. We define $\psi : \mathbb{R} \to \mathbb{R}$ by

$$\psi(r) := \begin{cases} (1-\tau) r^2 \,, & \text{if} \quad r < 0 \,, \\ \tau r^2 \,, & \text{if} \quad r \geqslant 0 \,. \end{cases}$$

Clearly, ψ is convex and thus by (Steinwart and Christmann, 2008, Lemma A.6.5) ψ is locally Lipschitz continuous. Moreover, for $y \in [-M, M]$ (see Steinwart and Christmann, 2008, Lemma A.6.8) we obtain

$$|L)|_{1,M} = \sup_{y \in [-M,M]} |\psi(y - \cdot)|_{1,M} \,,$$

$$
\begin{aligned}
&= \sup_{y\in[-M,M]} \sup_{t\in[-M,M]} |\psi'(y-t)|\,, \\
&= \max\{\tau, 1-\tau\} \sup_{y\in[-M,M]} \sup_{t\in[M,-M]} |2(y-t)|\,, \\
&= C_\tau\, 4M\,,
\end{aligned}
$$

where $C_\tau := \max\{\tau, 1-\tau\}$. ■

Note that L_τ being locally Lipschitz continuous implies by Steinwart and Christmann (2008, Lemma 2.19) that the corresponding risk $\mathcal{R}_{L_\tau,\mathrm{P}}(f)$ is also locally Lipschitz continuous. In addition, since $\mathcal{R}_{L_\tau,\mathrm{P}}(0) < \infty$, local Lipschitz continuity of L_τ also implies that L_τ is a *Nemitski loss* of order 2 in the sense of Definition 2.5.

3.2.4 Self-Calibration Inequalities

The risk $\mathcal{R}_{L_\tau,\mathrm{P}}(f)$ defined by (3.3) can be computed by treating the *inner* and the *outer* integrals separately. The inner integral leads to *inner L_τ-risks* which are key ingredients for establishing the self-calibration inequalities in Theorem 3.3. By Definition 2.6, for a distribution Q on $Y \subset \mathbb{R}$ and the L_τ-loss, the inner L_τ-risks are defined by

$$
\mathcal{C}_{L_\tau,Q}(t) := \int_Y L_\tau(y,t)dQ(y)\,, \qquad t \in \mathbb{R}\,, \tag{3.5}
$$

and the *minimal inner L_τ-risk* is defined by

$$
\mathcal{C}^*_{L_\tau,Q} := \inf_{t\in\mathbb{R}} \mathcal{C}_{L_\tau,Q}(t)\,. \tag{3.6}
$$

Here, the *inner risks* $\mathcal{C}_{L_\tau,Q}(\cdot)$ for a suitable classes of distributions Q on Y are considered as a template for $\mathcal{C}_{L_\tau,\mathrm{P}(\cdot|x)}(\cdot)$. Since L_τ is a convex loss function, the L_τ-inner risk is convex too. This is also illustrated in Figure 3.4 by considering different conditional distributions.

With the help of (3.5) and (3.6) one can immediately obtain risks (3.3) and the optimal risk (3.4) respectively, that is

$$
\mathcal{R}_{L_\tau,\mathrm{P}}(f) = \int_X \mathcal{C}_{L_\tau,\mathrm{P}(\cdot|x)}\, f(x)\, d\mathrm{P}_X(x)\,,
$$

and

$$
\mathcal{R}^*_{L_\tau,\mathrm{P}} = \int_X \mathcal{C}^*_{L_\tau,\mathrm{P}(\cdot|x)} d\mathrm{P}_X(x)\,.
$$

Furthermore, the *excess L_τ-risk*, when $\mathcal{R}^*_{L_\tau,\mathrm{P}} < \infty$ holds, is obtained by

$$
\mathcal{R}_{L_\tau,\mathrm{P}}(f) - \mathcal{R}^*_{L_\tau,\mathrm{P}} = \int_X \mathcal{C}_{L_\tau,\mathrm{P}(\cdot|x)}(f(x)) - \mathcal{C}^*_{L_\tau,\mathrm{P}(\cdot|x)} d\mathrm{P}_X(x)\,. \tag{3.7}
$$

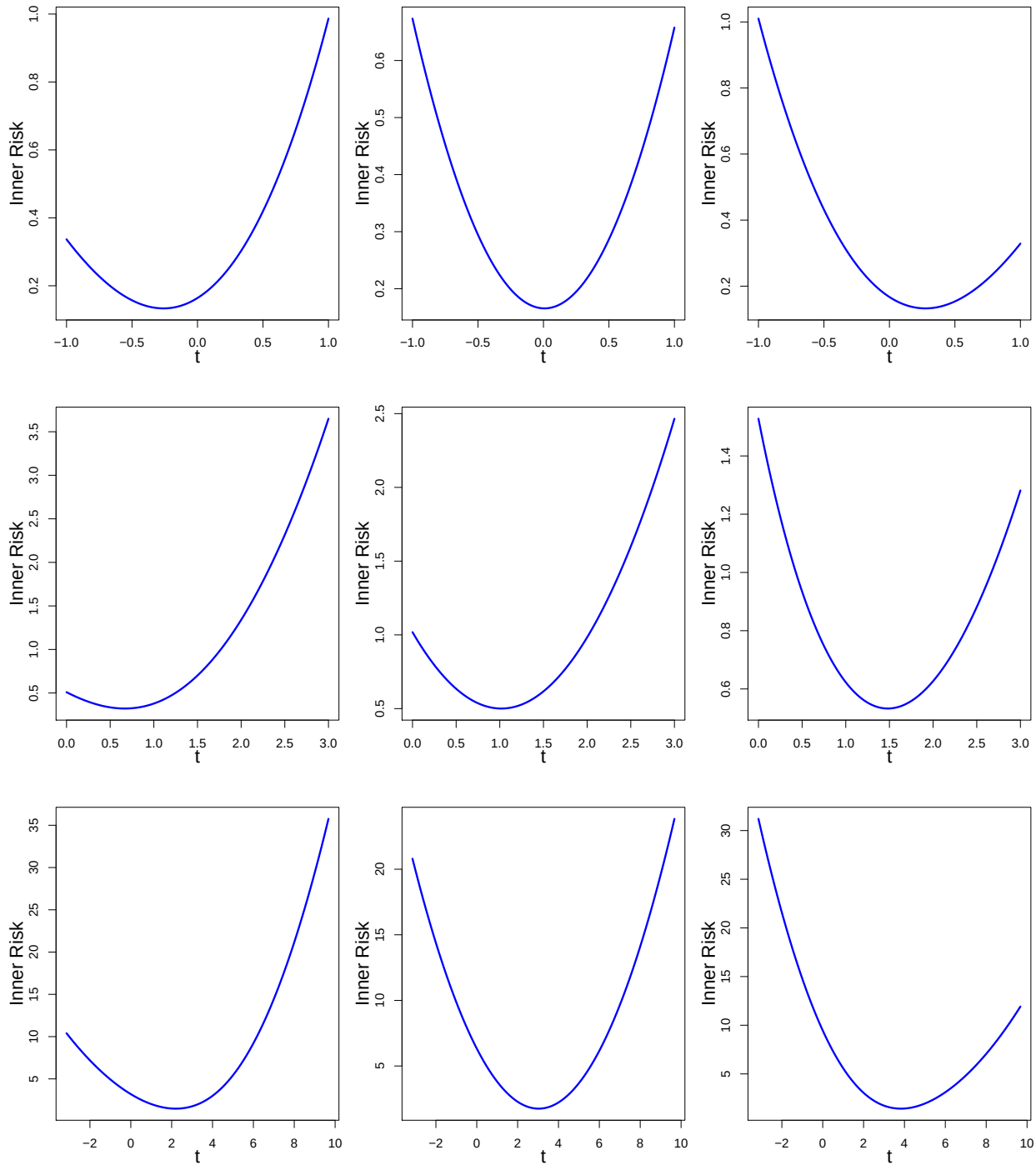

Figure 3.4: The L_τ-inner risks for $\tau = 0.25$ (left), $\tau = 0.50$ (middle) and $\tau = 0.75$ (right) assuming that the conditional distribution is a uniform distribution $U(-1, 1)$ (top), an exponential distirubiton $\mathsf{exp}(\lambda = 1)$(middle) and a mixture of normal distirubitons $0.4\mathcal{N}(0, 1) + 0.6\mathcal{N}(5, 3)$(bottom).

It is interesting to note that bounds on excess L_τ-risks can easily be translated from bounds on excess inner L_τ-risks. Based on this idea, we establish lower and upper bounds of excess inner

L_τ-risks in the following theorem.

Theorem 3.2. *Let $L_\tau : Y \times \mathbb{R} \to [0,\infty)$ be the ALS loss and Q be a distribution on $\mathbb{R}$ for which $|Q|_1 < \infty$ and $\mathcal{C}^*_{L_\tau,Q} < \infty$ hold. Then, for all $t \in \mathbb{R}$ and $\tau \in (0,1)$ we have*

$$c_\tau (t - t^*)^2 \le \mathcal{C}_{L_\tau,Q}(t) - \mathcal{C}^*_{L_\tau,Q} \le C_\tau (t - t^*)^2 \,, \tag{3.8}$$

where $c_\tau := \min\{\tau, 1-\tau\}$ and C_τ is defined in Lemma 3.1.

Proof of Theorem 3.2. Let us fix $\tau \in (0,1)$. Since the distirubiton Q on $\mathbb{R}$ has a finite first moment, i.e. $|Q|_1 < \infty$ holds, we obtain by (1.2) the τ-expectile t^* as a unique solution to

$$\tau \int_{y \ge t^*} (y - t^*) dQ(y) = (1-\tau) \int_{y < t^*} (t^* - y) dQ(y) \,, \tag{3.9}$$

For computing the excess inner risks of L_τ with respect to Q, we fix a $t \ge t^*$. Then we have

$$\begin{aligned}
&\int_{y<t} (y-t)^2 dQ(y) \\
&= \int_{y<t} (y - t^* + t^* - t)^2 dQ(y) \\
&= \int_{y<t} (y-t^*)^2 dQ(y) + 2(t^* - t) \int_{y<t} (y - t^*) dQ(y) + (t^* - t)^2 Q((-\infty, t)) \\
&= \int_{y<t^*} (y-t^*)^2 dQ(y) + \int_{t^* \le y < t} (y-t^*)^2 dQ(y) + (t^* - t)^2 Q((-\infty, t)) \\
&\quad + 2(t^* - t) \int_{y<t^*} (y - t^*) dQ(y) + 2(t^* - t) \int_{t^* \le y < t} (y - t^*) dQ(y) \,,
\end{aligned}$$

and

$$\begin{aligned}
&\int_{y \ge t} (y-t)^2 dQ(y) \\
&= \int_{y \ge t^*} (y-t^*)^2 dQ(y) - \int_{t^* \le y < t} (y-t^*)^2 dQ(y) + (t^* - t)^2 Q([t, \infty)) \\
&\quad + 2(t^* - t) \int_{y \ge t^*} (y - t^*) dQ(y) - 2(t^* - t) \int_{t^* \le y < t} (y - t^*) dQ(y) \,.
\end{aligned}$$

By using (3.5), (3.6) and (3.9) we obtain

$$\begin{aligned}
\mathcal{C}_{L_\tau,Q}(t) &= (1-\tau) \int_{y<t} (y-t)^2 dQ(y) + \tau \int_{y \ge t} (y-t)^2 dQ(y) \\
&= \tau \int_{y<t^*} (y-t^*)^2 dQ(y) + (1-\tau) \int_{y \ge t^*} (y-t^*)^2 dQ(y) \\
&\quad + 2(t^* - t) \Big(\tau \int_{y<t^*} (y - t^*) dQ(y) + (1-\tau) \int_{y \ge t^*} (y - t^*) dQ(y) \Big) \\
&\quad + (1-2\tau) \int_{t^* \le y < t} (y-t^*)^2 dQ(y) + 2(1-2\tau) \int_{t^* \le y < t} (y - t^*) dQ(y) \\
&\quad + (t^* - t)^2 (1-\tau) Q((-\infty, t)) + (t^* - t)^2 \tau Q([t, \infty))
\end{aligned}$$

$$= \mathcal{C}_{L_\tau,Q}(t^*) + (t^*-t)^2(1-\tau)Q((-\infty,t)) + (t^*-t)^2\tau Q([t,\infty)) + (1-2\tau)\int_{t^*\le y<t}(y-t^*)^2 + 2(t^*-t)(y-t^*)dQ(y)\,,$$

and this leads to the following excess inner L_τ-risk

$$\begin{aligned}
&\mathcal{C}_{L_\tau,Q}(t) - \mathcal{C}_{L_\tau,Q}(t^*)\\
&= (t^*-t)^2(1-\tau)Q((-\infty,t^*)) + (t^*-t)^2(1-\tau)Q([t^*,t)) + (t^*-t)^2\tau Q([t,\infty))\\
&\quad + (1-2\tau)\int_{t^*\le y<t}(y-t^*)^2 + 2(t^*-t)(y-t^*)dQ(y)\\
&= (t^*-t)^2\Big((1-\tau)Q((-\infty,t^*)) + \tau Q([t,\infty))\Big)\\
&\quad - \tau\int_{t^*\le y<t}(y-t^*)^2 + 2(t^*-t)(y-t^*)dQ(y)\\
&\quad + (t^*-t)^2(1-\tau)Q([t^*,t)) + (1-\tau)\int_{t^*\le y<t}(y-t^*)^2 + 2(t^*-t)(y-t^*)dQ(y)\\
&= (t^*-t)^2\Big((1-\tau)Q((-\infty,t^*)) + \tau Q([t,\infty))\Big) - \tau\int_{t^*\le y<t}(y-t^*)(y+t^*-2t)dQ(y)\\
&\quad + (1-\tau)\int_{t^*\le y<t}(y-t^*)^2 + 2(t^*-t)(y-t^*) + (t^*-t)^2dQ(y)\\
&= (t^*-t)^2\Big((1-\tau)Q((-\infty,t^*)) + \tau Q([t,\infty))\Big) + \tau\int_{t^*\le y<t}(y-t^*)(2t-t^*-y)dQ(y)\\
&\quad + (1-\tau)\int_{t^*\le y<t}(y-t)^2dQ(y)\,.
\end{aligned} \tag{3.10}$$

Let us define $c_\tau := \min\{\tau, 1-\tau\}$, then (3.10) leads to the following lower bound of excess inner L_τ-risk when $t \ge t^*$:

$$\begin{aligned}
&\mathcal{C}_{L_\tau,Q}(t) - \mathcal{C}_{L_\tau,Q}(t^*)\\
&\ge c_\tau(t^*-t)^2\Big(Q((-\infty,t^*)) + Q([t,\infty))\Big) + c_\tau\int_{t^*\le y<t}(y-t^*)(2t-t^*-y) + (y-t)^2dQ(y)\\
&= c_\tau(t^*-t)^2\Big(Q((-\infty,t^*)) + Q([t,\infty))\Big) + c_\tau\int_{t^*\le y<t}(t^*)^2 + 2tt^* + t^2dQ(y)\\
&= c_\tau(t^*-t)^2\Big(Q((-\infty,t^*)) + Q([t,\infty))\Big) + c_\tau(t^*-t)^2Q([t^*,t))\\
&= c_\tau(t^*-t)^2\,.
\end{aligned} \tag{3.11}$$

Likewise, the excess inner L_τ-risk when $t < t^*$ is

$$\begin{aligned}
&\mathcal{C}_{L_\tau,Q}(t) - \mathcal{C}_{L_\tau,Q}(t^*)\\
&= (t^*-t)^2\Big((1-\tau)Q((-\infty,t) + \tau)Q([t^*,\infty))\Big) + \tau\int_{t\le y<t^*}(y-t)^2dQ(y)\\
&\quad + (1-\tau)\int_{t\le y<t^*}(t^*-y)(y+t^*-2t)dQ(y)\,,
\end{aligned} \tag{3.12}$$

that also leads to the lower bound (3.11). Now, for the proof of upper bound of the excess inner L_τ-risks, we define $C_\tau := \max\{\tau, 1-\tau\}$. Then (3.10) leads to the following upper bound

of excess inner L_τ-risks when $t \geq t^*$:

$$\begin{aligned}\mathcal{C}_{L_\tau,Q}(t) - \mathcal{C}_{L_\tau,Q}(t^*) &\leq C_\tau(t^* - t)^2\Big(Q((-\infty, t^*)) + Q([t,\infty))\Big) \\ &\quad + C_\tau \int_{t^* \leq y < t} \big((y - t^*)(2t - t^* - y) + (y - t)^2\big) dQ(y) \\ &= C_\tau(t^* - t)^2 \,. \end{aligned} \tag{3.13}$$

Analogously, for the case of $t < t^*$, (3.12) also leads to the upper bound (3.13) for excess inner L_τ-risks. ■

Recall that the empirical methods of estimating expectile using the L_τ-loss typically lead to a function f_D for which $\mathcal{R}_{L_\tau,\mathrm{P}}(f_D)$ is close to $\mathcal{R}^*_{L_\tau,\mathrm{P}}$ with high probability. The convexity of L_τ then ensures that f_D approximates $f^*_{L_\tau,\mathrm{P}}$ in a weak sense, that is, in probability P_X, see (Steinwart, 2007, Remark 3.18). However, no guarantee on the speed of convergence can be given. The following theorem addresses this issue by establishing a so-called calibration inequality for the excess L_τ-risk.

Theorem 3.3. *Let L_τ be the ALS loss function and P be a distribution on $\mathbb{R}$ satisfying $|\mathrm{P}|_2 < \infty$. Moreover, assume that for any $\tau \in (0,1)$, the conditional τ-expectile $f^*_{L_\tau,\mathrm{P}}(x) < \infty$ holds and $f^*_{L_\tau,\mathrm{P}} \in L_2(\mathrm{P}_X)$ for P_X-almost all $x \in X$. Then, for all measurable $f : X \to \mathbb{R}$, we have*

$$C_\tau^{-1/2}(\mathcal{R}_{L_\tau,\mathrm{P}}(f) - \mathcal{R}^*_{L_\tau,\mathrm{P}})^{1/2} \leq \|f - f^*_{L_\tau,\mathrm{P}}\|_{L_2(P_X)} \leq c_\tau^{-1/2}\,(\mathcal{R}_{L_\tau,\mathrm{P}}(f) - \mathcal{R}^*_{L_\tau,\mathrm{P}})^{1/2}\,, \tag{3.14}$$

where $c_\tau := \min\{\tau, 1-\tau\}$ and C_τ is defined in Lemma 3.1.

Proof of Theorem 3.3. For a fixed $x \in X$, we write $t := f(x)$ and $t^* := f^*_{L_\tau,\mathrm{P}}(x)$. By Theorem 3.2, for $Q := \mathrm{P}(\,\cdot\,|x)$, we then immediately obtain

$$C_\tau^{-1}(\mathcal{C}_{L_\tau,\mathrm{P}(\cdot|x)}(f(x)) - \mathcal{C}^*_{L_\tau,\mathrm{P}(\cdot|x)}) \leq |f(x) - f^*_{L_\tau,\mathrm{P}}(x)|^2 \leq c_\tau^{-1}\,(\mathcal{C}_{L_\tau,\mathrm{P}(\cdot|x)}(f(x)) - \mathcal{C}^*_{L_\tau,\mathrm{P}(\cdot|x)})$$

Integrating with respect to P_X leads to the assertion. ■

Note that the right-hand side of (3.14) in particular ensures that $f_D \to f^*_{L_\tau,\mathrm{P}}$ in $L_2(P_X)$ whenever $\mathcal{R}_{L_\tau,\mathrm{P}}(f_D) \to \mathcal{R}^*_{L_\tau,\mathrm{P}}$. In addition, the convergence rates can be directly translated. Moreover, the inequality on the left shows that modulo constants the calibration inequality is sharp. We will also use this left inequality when bounding the approximation error function for Gaussian RBF kernels in the proof of Theorem 4.3. Finally, in Lemma 4.12, we will show that the inequality (3.14) leads to the bound of the approximation error in case of generic kernels if and only if the target function is in a real interpolation space, that is, $f^*_{L_\tau,\mathrm{P}} \in [L_2(P_X), H]_{\beta,\infty}$ for some $\beta \in (0,1)$.

3.2.5 Supremum and Variance Bounds

Like the calibration inequality of Theorem 3.3, supremum and variance bounds for the L_τ-loss are also useful for analyzing the statistical properties of any L_τ-based empirical risk minimization scheme as we will see in Chapter 4 while establishing oracle inequalities for the SVM-type learning algorithm (1.7). In the following Lemma we present the supremum and variance bounds for the L_τ-loss.

Lemma 3.4. *Let $X \subset \mathbb{R}^d$ be a non-empty set, $Y \subset [-M, M]$ be a closed subset where $M > 0$, and P be a distribution on $X \times Y$. Additionally, we assume that $L_\tau : Y \times \mathbb{R} \to [0, \infty)$ is the ALS loss and $f^*_{L_\tau,\mathrm{P}}$ is the conditional τ-expectile for fixed $\tau \in (0,1)$. Then for all measurable $f : X \to [-M, M]$ we have*

i) $\|L_\tau \circ f - L_\tau \circ f^*_{L_\tau,\mathrm{P}}\|_\infty \le 4\, C_\tau\, M^2\,.$

ii) $\mathbb{E}_P(L_\tau \circ f - L_\tau \circ f^*_{L_\tau,\mathrm{P}})^2 \le 16\, C_\tau^2\, c_\tau^{-1}\, M^2(\mathcal{R}_{L_\tau,\mathrm{P}}(f) - \mathcal{R}^*_{L_\tau,\mathrm{P}})\,.$

Proof of Lemma 3.4. i) Since L_τ can be clipped at M and the conditional τ-expectile satisfies $f^*_{L_\tau,\mathrm{P}}(x) \in [-M, M]$ almost surely for all $x \in X$, we obtain

$$\begin{aligned} |L_\tau(y, f(x)) - L_\tau(y, f^*_{L_\tau,\mathrm{P}}(x))| &\le \max\{\tau, 1-\tau\} \sup_{y,t \in [-M,M]} (y-t)^2 \\ &= C_\tau\, 4M^2\,, \end{aligned}$$

which holds for all $f : X \to [-M, M]$ and all $(x, y) \in X \times Y$.

ii) Using local Lipschitz continuity of L_τ and Theorem 3.3, we obtain

$$\begin{aligned} \mathbb{E}_\mathrm{P}(L_\tau \circ f - L_\tau \circ f^*_{L_\tau,\mathrm{P}})^2 &\le |L_\tau|^2_{1,M}\ \mathbb{E}_{\mathrm{P}_X}|f - f^*_{L_\tau,\mathrm{P}}|^2 \\ &\le 16 c_\tau^{-1} C_\tau^2\, M^2\, (\mathcal{R}_{L_\tau,\mathrm{P}}(f) - \mathcal{R}^*_{L_\tau,\mathrm{P}})\,. \end{aligned}$$

■

Chapter 4

Learning Rates for Kernel-Based Expectile Regression

We devote this chapter for statistical analysis of SVM-type learning algorithm for expectile regression. This includes establishing learning rates considering Gaussian RBF kernels, see Section 4.1, and generic kernels shown in Section 4.2. For $\tau = 0.5$, we also compare our achieved learning rates with the best-known learning rates for least squares regression in both scenarios.

Let us assume that P is the probability distribution on $X \times Y$, where $X \subset \mathbb{R}^d$ and $Y \subset \mathbb{R}$. In addition we assume that H is a separable RKHS over X associated to a measurable and positive definite kernel k. Furthermore, we assume that k is bounded that is, $\|k\|_\infty := \sup_{x \in X} \sqrt{k(x,x)} < \infty$ which implies that H consists of bounded functions with $\|f\|_\infty \leq \|k\|_\infty \|f\|_H$ for all $f \in H$. Moreover, we denote by H_γ a Gaussian RKHS associated to a Gaussian RBF kernel k_γ defined by (2.7), where γ is called the width parameter and it is usually determine in a data-dependent way, e.g. by cross validation.

We now recall that given an i.i.d. data set $D := ((x_1, y_1), \ldots, (x_n, y_n))$ drawn from some unknown distribution P, SVMs construct a predictor $f_{\mathrm{D},\lambda}$ by solving the convex optimization problem of the form

$$f_{\mathrm{D},\lambda} = \arg\min_{f \in H} \lambda \|f\|_H^2 + \mathcal{R}_{L_\tau,\mathrm{D}}(f)\,, \tag{4.1}$$

where $\lambda > 0$ is a regularization parameter and $\mathcal{R}_{L_\tau,\mathrm{D}}(\cdot)$ is the empirical L_τ-risk defined by (1.6). A typical way to access the quality of an estimator f_D ($f_{\mathrm{D},\lambda}$ in our case) produced by a *learning method* like (4.1), is to measure its distance to the target function $f^*_{L_\tau,\mathrm{P}}$, e.g. in terms of $\|f_\mathrm{D} - f^*_{L_\tau,\mathrm{P}}\|_{L_2(\mathrm{P}_X)}$. For estimators obtained by some empirical risk minimization scheme, however, one can hardly ever estimate this L_2-norm directly. Instead, the standard approach

of statistical learning theory deals with the excess risk $\mathcal{R}_{L_\tau,\mathrm{P}}(f_\mathrm{D}) - \mathcal{R}^*_{L_\tau,\mathrm{P}}$. In other words, one is interested to know the probability that $\mathcal{R}_{L_\tau,\mathrm{P}}(f_\mathrm{D})$ is close to $\mathcal{R}^*_{L_\tau,\mathrm{P}}$. This leads to the notion of *consistency* of learning methods. In the following theorem, we show that our considered learning method (4.1) for estimating the conditional τ-expectiles is universally consistent for a suitably chosen data-dependent null sequence of regularization parameters (λ_n) with $\lambda_n > 0$.

Theorem 4.1. *Let* P *be a distribution on* $X \times Y$ *with* $|\mathrm{P}|_2 < \infty$, L_τ *be the ALS loss, and* $f^*_{L_\tau,\mathrm{P}}$ *be the conditional* τ*-expectile function. Moreover, let* k *be a bounded, measurable kernel whose RKHS is separable and dense in* $L_2(\mathrm{P}_X)$. *Then for all sequences* $\lambda_n \to 0$ *with* $\lambda_n^4 n \to \infty$ *and all* $\varepsilon > 0$, *we have*

$$\lim_{n\to\infty} P^n\big(D \in (X \times \mathbb{R})^n : \mathcal{R}_{L_\tau,\mathrm{P}}(f_{\mathrm{D},\lambda_n}) - \mathcal{R}^*_{L_\tau,\mathrm{P}} > \varepsilon\big) = 0\,, \tag{4.2}$$

and

$$\lim_{n\to\infty} P^n\big(D \in (X \times \mathbb{R})^n : \|f_{\mathrm{D},\lambda_n} - f^*_{L_\tau,\mathrm{P}}\|_0 > \varepsilon\big) = 0\,,$$

where $\|g\|_0 := \int \min\{1,|g|\} d\mathrm{P}_X$ *is a translation-invariant metric describing convergence in probability* P_X. *Moreover, if* $f^*_{L_\tau,\mathrm{P}} \in L_2(\mathrm{P}_X)$, *then*

$$\lim_{n\to\infty} P^n\big(D \in (X \times \mathbb{R})^n : \|f_{\mathrm{D},\lambda_n} - f^*_{L_\tau,\mathrm{P}}\|_{L_2(\mathrm{P}_X)} > \varepsilon\big) = 0\,.$$

Proof of Theorem 4.1. The first convergence follows from (Steinwart and Christmann, 2008, Theorem 9.1) and the second convergence is a consequence of the first convergence and (Steinwart and Christmann, 2008, Corollary 3.62), where we note that we do not need the completeness of X since we already know the existence and uniqueness of $f^*_{L_\tau,\mathrm{P}}$. Finally, the third convergence is a consequence of the first convergence and the self-calibration inequality in Theorem 3.3. ■

Theorem 4.1 shows that, for sufficiently large training sets, the learning method (4.1) produces nearly optimal decision function with high probability, without knowing any specifics of the data-generating distribution P. However, note that the notion of (universal) consistency is purely of asymptotic nature. In other words, (4.2) does not quantify the speed of convergence of $\mathcal{R}_{L_\tau,\mathrm{P}}(f_{\mathrm{D},\lambda_n})$ to $\mathcal{R}^*_{L_\tau,\mathrm{P}}$, which is a more natural and practical requirement. We cope with this issue in Section 4.1 and Section 4.2 by establishing learning rates for learning method (4.1) using Gaussian RBF kernels and generic kernels, respectively. However, learning with guaranteed rates of convergence almost always requires assumptions on the unknown distribution P. In fact, by *no-free-lunch* theorem (see Devroye et al, 1996, Theorem 7.2) no method learns with a fixed rate and confidence for all distributions P. In other words, no learning method enjoys a uniform learning rates, see e.g. Steinwart and Christmann (2008, Corollary 6.7 & 6.8).

4.1 Learning Rates Assuming Gaussian RBF Kernels

This section presents learning rates for SVM-type learning algorithm for the expectile regression considering Gaussian RBF kernels. In Section 4.1.1, we first establish an improved entropy number bound for the Gaussian RKHSs to control its capacity. Section 4.1.2 gives an upper bound for an approximation error function. Then we use the results of Section 4.1.1 and Section 4.1.2 together already established results in Chapter 3 and establish learning rates in Section 4.1.3 under the assumptions that $Y \subseteq [-M, M]$ where $M > 0$ and the target function $f^*_{L_\tau,\mathrm{P}}$ is smooth in a Besov sense. In Section 4.1.4, we use the data-dependent parameters' selection method and obtain the same learning rates adaptively, that is, without knowing the smoothness parameter. Finally, we replace the assumption of bounded regression with the assumption of exponential decay of Y-tails in Section 4.1.5 and obtain learning rates under this scenario.

4.1.1 Improved Entroy Bounds for the Gaussian RKHSs

In order to control the capacity of a RKHS in terms of some bounds, one way is to estimate eigenvalues of a linear operator induced by the kernel k. Recall that given a kernel k over X with a separable RKHS H and a distribution P_X, the integral operator $T_k : L_2(\mu) \to L_2(\mu)$ is defined by

$$T_k f(\cdot) := \int_X k(x, \cdot) f(x) d\mu(x) \,,$$

The operator T_k is compact, positive, self-adjoint and nuclear (see Steinwart and Christmann, 2008, Theorem 4.27), and has at most countably many non-zero (and non-negative) eigenvalues $\lambda_i(T_k)$. Ordering these eigenvalues (with geometric multiplicities) and extending the corresponding sequence by zeros, if there are only finitely many non-zero eigenvalues, we obtain the *extended sequence of eigenvalues* $(\lambda_i(T_k))_{i\geq 1}$ that satisfies $\sum_{i=1}^{\infty} \lambda_i(T_k) < \infty$ (see Steinwart and Christmann, 2008, Theorem 7.29). This summability implies that for some constant $a > 1$ and $i \geq 1$, we have $\lambda_i(T_k) \leq a i^{-1}$. By Steinwart et al (2009), this eigenvalues assumption can converge even faster to zero, that is, for $p \in (0, 1)$, we often have

$$\lambda_i(T_k) \leq a\, i^{-\frac{1}{p}}, \qquad i \geq 1. \tag{4.3}$$

It turns out that the speed of convergence of $\lambda_i(T_k)$ influences learning rates for SVMs. For instance, Blanchard et al (2008) used (4.3) to establish learning rates for SVMs using hinge loss, and Caponnetto and De Vito (2007) and Mendelson and Neeman (2010) for SVMs using least squares loss.

Another way to control the capacity of RKHS H is based on *entropy numbers* that is the dual of covering numbers. Let $T : E \to F$ be a bounded, linear operator between the Banach spaces E and F, and $i \geq 1$ be an integer. Then the i-th (dyadic) entropy number of T is defined by

$$e_i(T) := \inf \left\{ \epsilon > 0 : \exists x_1, \ldots, x_{2^{i-1}} \text{ such that } TB_E \subset \cup_{j=1}^{2^{i-1}} (x_j + \epsilon B_F) \right\} ,$$

where B_F is called closed unit ball of F, see Definition 2.18. In the Hilbert space case, the eigenvalues and entropy number decay are closely related. For example, Steinwart (2009) showed that (4.3) is equivalent (modulo a constant only depending on p) to

$$e_i(\mathrm{id} : H \to L_2(P_X)) \leq \sqrt{a}\, i^{-\frac{1}{2p}} , \qquad i \geq 1 , \tag{4.4}$$

It is further shown by Steinwart (2009) that (4.4) implies a bound on average entropy numbers, that is, for empirical distribution associated to the data set $D_X := (x_1, \cdots, x_n) \in X^n$, the average entropy number, modulo some constant, is

$$\mathbb{E}_{D_X \sim P_X^n} e_i(\mathrm{id} : H \to L_2(P_X)) \leq a\, i^{-\frac{1}{2p}}, \qquad i \geq 1 ,$$

which is used in Theorem 2.20 to establish the general oracle inequality for SVMs. A bound of the form (4.4) was also derived in (Steinwart and Christmann, 2008, Theorem 6.27) for Gaussian RBF kernels. To be more precise, let $X \subset \mathbb{R}^d$ be a closed unit Euclidean ball. Then for all $\gamma \in (0, 1]$ and $p \in (0, 1)$, there exists a constant $c_{p,d}(X)$ such that

$$e_i(\mathrm{id} : H_\gamma(X) \to l_\infty(X)) \leq c_{p,d}(X)\, \gamma^{-\frac{d}{p}}\, i^{-\frac{1}{p}} , \qquad , i \geq 1 , \tag{4.5}$$

which has been used by Eberts and Steinwart (2013) to establish leaning rates for least squares SVMs. Note that the constant $c_{p,d}(X)$ depends on p in an unknown manner. To address this issue, we use the result of van der Vaart and van Zanten (2009, Lemma 4.5) and derive an improved entropy number bound in the following theorem. Our improved entropy bound provides an upper bound for $c_{p,d}(X)$ whose dependence on p is known explicitly. In addition, we will further see in Corollary 4.7 that by using this improved bound, we establish up to logarithmic factor minimax optimal learning rates for L_τ.

Theorem 4.2. *Let $X \subseteq \mathbb{R}^d$ be the closed unit Euclidean ball. Then there exists a constant $K > 0$, such that, for all $p \in (0, 1)$, $\gamma \in (0, 1]$ and $i \geq 1$, we have*

$$e_i(\mathrm{id} : H_\gamma(X) \to l_\infty(X)) \leq (3K)^{\frac{1}{p}} \left(\frac{d+1}{ep} \right)^{\frac{d+1}{p}} \gamma^{-\frac{d}{p}}\, i^{-\frac{1}{p}} \tag{4.6}$$

Proof of Theorem 4.2. By van der Vaart and van Zanten (2009, Lemma 4.5), the $\|\cdot\|_\infty$-log covering numbers of unit ball $B_\gamma(X)$ of the Gaussian RKHS $H_\gamma(X)$ for all $\gamma \in (0,1)$ and $\varepsilon \in (0, \frac{1}{2})$ satisfy

$$\ln \mathcal{N}(B_\gamma(X), \|\cdot\|_\infty, \varepsilon) \leq K_d \left(\log \frac{1}{\varepsilon}\right)^{d+1} \gamma^{-d}, \tag{4.7}$$

where $K_d \geq 1$ is a constant depending only on d. From this, obtain

$$\sup_{\varepsilon \in (0,\frac{1}{2})} \varepsilon^p \ln \mathcal{N}(B_\gamma(X), \|\cdot\|_\infty, \varepsilon) \leq K_d \gamma^{-d} \sup_{\varepsilon \in (0,\frac{1}{2})} \varepsilon^p \left(\log \frac{1}{\varepsilon}\right)^{d+1}.$$

Let $h(\varepsilon) := \varepsilon^p \left(\log \frac{1}{\varepsilon}\right)^{d+1}$. In order to obtain the optimal value of $h(\varepsilon)$, we differentiate it with respect to ε

$$\frac{dh(\varepsilon)}{d\varepsilon} = p\, \varepsilon^{p-1} \left(\log \frac{1}{\varepsilon}\right)^{d+1} - \varepsilon^p (d+1) \left(\log \frac{1}{\varepsilon}\right)^d \frac{1}{\varepsilon},$$

and by setting $\frac{dh(\varepsilon)}{d\varepsilon} = 0$ we obtain $\log \frac{1}{\varepsilon} = \frac{d+1}{p}$ which finally yields

$$\varepsilon^* = \frac{1}{e^{\frac{d+1}{p}}}.$$

By plugging ε^* into $h(\varepsilon)$, we obtain

$$h(\varepsilon^*) = \left(\frac{d+1}{ep}\right)^{d+1},$$

and consequently, the $\|\cdot\|_\infty$-log covering numbers (4.7) satisfy

$$\ln \mathcal{N}(B_\gamma(X), \|\cdot\|_\infty, \varepsilon) \leq K_d \left(\frac{d+1}{ep}\right)^{d+1} \gamma^{-d} \varepsilon^{-p} = \left(\frac{a^{\frac{1}{p}}}{\varepsilon}\right)^p,$$

where $a := K_d \left(\frac{d+1}{ep}\right)^{d+1} \gamma^{-d}$. Now, by Lemma 2.19 the bound on entropy number of the Gaussian RBF kernel is

$$e_i(\mathrm{id} : \mathcal{H}_\gamma(X) \to l_\infty(X)) \leq (3a)^{\frac{1}{p}} i^{-\frac{1}{p}} = (3K_d)^{\frac{1}{p}} \left(\frac{d+1}{ep}\right)^{\frac{d+1}{p}} \gamma^{-\frac{d}{p}} i^{-\frac{1}{p}},$$

for all $i \geq 1$, $\gamma \in (0,1)$. ■

4.1.2 Approximation Error Bounds

Given a loss function L_τ, a distribution P with $\mathcal{R}^*_{L_\tau,\mathrm{P}} < \infty$ and the Gaussian RKHS H_γ, we define for all $\lambda > 0$ the approximation error function $\mathcal{A}_\gamma : [0,\infty) \to [0,\infty)$ by

$$\mathcal{A}_\gamma(\lambda) := \inf_{f \in H_\gamma} \lambda \|f\|^2_{H_\gamma} + \mathcal{R}_{L_\tau,\mathrm{P}}(f) - \mathcal{R}^*_{L_\tau,\mathrm{P}}, \qquad \gamma > 0, \lambda \geq 0. \tag{4.8}$$

For $\lambda > 0$ and $\gamma > 0$, the approximation error function $\mathcal{A}_\gamma(\lambda)$ quantifies how well risk of an infinite sample SVM with RKHS H_γ, that is, $\lambda\|f\|_{H_\gamma}^2 + \mathcal{R}_{L_\tau,\mathrm{P}}(f)$ approximates the optimal risk $\mathcal{R}^*_{L_\tau,\mathrm{P}}$.

In order to bound $\mathcal{A}_\gamma(\lambda)$, we first need to know one important feature of the target function $f^*_{L_\tau,\mathrm{P}}$, namely, the *regularity* which, roughly speaking, measures the smoothness of the target function. Different function spaces norms e.g. Hölder norms, Besov norms or Triebel-Lizorkin norms can be used to capture this regularity. In this work, following Eberts and Steinwart (2013) and Meister and Steinwart (2016), we assume that the target function $f^*_{L_\tau,\mathrm{P}}$ is in a Sobolev or a Besov space. Recall Tartar (2007, Definition 5.1) and Adams and Fournier (2003, Definition 3.1 and 3.2) that for any integer $k \geq 0$, $1 \leq p \leq \infty$ and a subset $\Omega \subset \mathbb{R}^d$ with non-empty interior, the Sobolev space $W_p^k(\Omega)$ of order k is defined by

$$W_p^k(\Omega) := \left\{ f \in L_p(\Omega) : D^{(\alpha)} f \in L_p(\Omega) \text{ exists for all } \alpha \in \mathbb{N}_0^d \text{ with } |\alpha| \leq k \right\},$$

with the norm

$$\|f\|_{W_p^k(\Omega)} := \begin{cases} \left(\sum_{|\alpha| \leq k} \|D^{(\alpha)} f\|_{L_p(\Omega)}^p \right)^{\frac{1}{p}}, & \text{if } \ p \in [1, \infty), \\ \max_{|\alpha| \leq k} \|D^{(\alpha)} f\|_{L_\infty(\Omega)}, & \text{if } \ p = \infty, \end{cases}$$

where $D^{(\alpha)}$ is the α-th weak partial derivative for the multi-index $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0^d$ of modulus $|\alpha| = |\alpha_1| + \cdots + |\alpha_d|$. In other words, the Sobolev space is the space of functions with sufficiently many weak derivatives and equipped with a norm that measures both the size and the regularity of the contained functions. Note that $W_p^k(\Omega)$ is a Banach space (Tartar, 2007, Lemma 5.2). Moreover, by Adams and Fournier (2003, Theorem 3.6), $W_p^k(\Omega)$ is separable if $p \in [1, \infty)$, and is uniformly convex and reflexive if $p \in (1, \infty)$. Furthermore, for $p = 2$, $W_2^k(\Omega)$ is a separable Hilbert space that we denote by $H_k(\Omega)$. Despite the aforementioned advantages, Sobolev spaces can not be immediately applied when k is non-integral or when $p < 1$, however, the smoothness spaces for these extended parameters are also needed when engaging nonlinear approximation. This shortcoming of Sobolev spaces is covered by Besov spaces which bring together all functions for which the modulus of smoothness have a common behavior. Let us first recall DeVore and Sharpley (1993, Section 2) and DeVore and Popov (1988, Section 2) that for a function $f : \Omega \to \mathbb{R}$ with $f \in L_p(\Omega)$ for $p \in (0, \infty]$ and $s \in \mathbb{N}$, the modulus of smoothness of order s of a function f is defined by

$$w_{s,L_p(\Omega)}(f, t) = \sup_{\|h\|_2 \leq t} \|\triangle_h^s(f, \cdot)\|_{L_p(\Omega)}, \qquad t \geq 0,$$

where the s-th difference $\triangle_h^s(f,\cdot)$ given by

$$\triangle_h^s(f,x) := \begin{cases} \sum_{i=0}^s \binom{s}{i}(-1)^{s-i} f(x+ih) & \text{if} \quad x, x+h, \ldots, x+sh \in \Omega\,, \\ 0, & \text{otherwise}\,, \end{cases}$$

for $h \in \mathbb{R}^d$ is used to measure the smoothness. Note that $w_{s,L_p(\Omega)}(f,t) \to 0$ as $t \to 0$, which means that the faster this convergence to 0 is, the smoother is the function f. For more details on properties of the modulus of smoothness, we refer the reader to Nikol'skii (2012, Chapter 4.2). Now for $0 < p, q \le \infty$, $\alpha > 0$, $s := \lfloor \alpha \rfloor + 1$, the Besov space $B^\alpha_{p,q}(\Omega)$ based on modulus of smoothness for domain $\Omega \subset \mathbb{R}^d$, see for instance DeVore (1998, Section 4.5), Nikol'skii (2012, Chapter 4.3) and DeVore and Sharpley (1993, Section 2), is defined by

$$B^\alpha_{p,q}(\Omega) := \{f \in L_p(\Omega) : |f|_{B^\alpha_{p,q}(\Omega)} < \infty\}\,,$$

where the semi-norm $|\cdot|_{B^\alpha_{p,q}(\Omega)}$ is defined by

$$|f|_{B^\alpha_{p,q}(\Omega)} := \Big(\int_0^\infty (t^{-\alpha} w_{s,L_p(\Omega)}(f,t))^q \frac{dt}{t} \Big)^{\frac{1}{q}}, \qquad q \in (0,\infty)\,,$$

and for $q = \infty$, the semi-norm is defined by

$$|f|_{B^\alpha_{p,q}(\Omega)} := \sup_{t>0} (t^{-\alpha} w_{s,L_p(\Omega)}(f,t))\,.$$

In other words, Besov spaces are collections of functions f with common smoothness. For more general definition of Besov-like spaces, we refer to Meister and Steinwart (2016, Section 4.1). Note that $\|f\|_{B^\alpha_{p,q}(\Omega)} := \|f\|_{L_p(\Omega)} + |f|_{B^\alpha_{p,q}(\Omega)}$ is a norm of $B^\alpha_{p,q}(\Omega)$, see e.g. DeVore and Sharpley (1993, Section 2) and DeVore and Popov (1988, Section 2). It is well known (see e.g. Nikol'skii, 2012, Section 4.1), that $W^\alpha_p(\Omega) \subset B^\alpha_{p,\infty}(\Omega)$ for all $1 \le p \le \infty$, $p \ne 2$, where for $p = q = 2$ the Besov space and the Sobolev space are isomorphic.

In the next step, we find a function $f_0 \in H_\gamma$ such that both the regularization term $\lambda \|f_0\|^2_{H_\gamma}$ and the excess risk $\mathcal{R}_{L_\tau,\mathrm{P}}(f_0) - \mathcal{R}^*_{L_\tau,\mathrm{P}}$ are small. For this, we define the function $K_\gamma : \mathbb{R}^d \to \mathbb{R}$ (see Eberts and Steinwart, 2013) by

$$K_\gamma(x) := \sum_{j=1}^r \binom{r}{j} (-1)^{1-j} \frac{1}{j^d} \Big(\frac{2}{\gamma^2 \pi}\Big)^{\frac{d}{2}} \exp\Big(-\frac{2\|x\|_2^2}{j^2\gamma^2}\Big)\,,$$

for all $r \in \mathbb{N}$, $\gamma > 0$ and $x \in \mathbb{R}^d$. Additionally, we assume that there exists a function $f^*_{L_\tau,\mathrm{P}} : \mathbb{R}^d \to \mathbb{R}$ satisfying $f^*_{L_\tau,\mathrm{P}} \in L_2(\mathbb{R}^d) \cap L_\infty(\mathbb{R}^d)$ and $\mathcal{R}_{L_\tau,\mathrm{P}}(f^*_{L_\tau,\mathrm{P}}) = \mathcal{R}^*_{L_\tau,\mathrm{P}}$. Then f_0 is defined by

$$f_0(x) := K_\gamma * f^*_{L_\tau,\mathrm{P}}(x) := \int_{\mathbb{R}} K_\gamma(x-t) f^*_{L_\tau,\mathrm{P}}(t)dt\,, \qquad x \in \mathbb{R}\,. \tag{4.9}$$

With these preparations, we now establish an upper bound for the approximate error function $\mathcal{A}_\gamma(\lambda)$.

Theorem 4.3. *Let $L_\tau : Y \times \mathbb{R} \to [0,\infty)$ be the ALS loss, P be a probability distribution on $\mathbb{R}^d \times Y$, and P_X be the marginal distribution of P on $\mathbb{R}^d$ such that $X := \operatorname{supp} \mathrm{P}_X$ satisfies $\mathrm{P}_X(\partial X) = 0$. Moreover, assume that the conditional τ-expectile $f^*_{L_\tau,\mathrm{P}}$ satisfies $f^*_{L_\tau,\mathrm{P}} \in L_2(\mathbb{R}^d) \cap L_\infty(\mathbb{R}^d)$ as well as $f^*_{L_\tau,\mathrm{P}} \in B^\alpha_{2,\infty}(\mathrm{P}_X)$ for some $\alpha \geq 1$. In addition, assume that k_γ is the Gaussian RBF kernel over X with associated RKHS H_γ. Then for $f_0 \in H_\gamma$ defined by (4.9), we have*

$$\|L_\tau \circ f_0\|_\infty \leq 4C_\tau(M + 2^s \|f^*_{L_\tau,\mathrm{P}}\|_{L_\infty(\mathbb{R}^d)})^2 =: B_0$$

Moreover, for all $\gamma \in (0,1]$ and $\lambda > 0$, we have

$$\|f_0\|^2_{H_\gamma} + \mathcal{R}_{L_\tau,\mathrm{P}}(f_0) - \mathcal{R}^*_{L_\tau,\mathrm{P}} \leq C_1 \lambda \gamma^{-d} + C_{\tau,s} \gamma^{2\alpha}\,, \tag{4.10}$$

where the constant $C_1 > 0$ and the constant $C_{\tau,s} > 0$ depends on s and τ.

Proof of Theorem 4.3. The assumption $f^*_{L_\tau,\mathrm{P}} \in L_2(\mathbb{R}^d)$ and (Eberts and Steinwart, 2013, Theorem 2.3) immediately yield that $f_0 := K_\gamma * f^*_{L_\tau,\mathrm{P}} \in H_\gamma$, i.e. f_0 is contained in H_γ. Furthermore, the latter theorem together with the assumption $f^*_{L_\tau,\mathrm{P}} \in L_\infty(\mathbb{R}^d)$ yields that for all $x \in X$

$$|K * f^*_{L_\tau,\mathrm{P}}(x)| \leq (2^s - 1)\|f^*_{L_\tau,\mathrm{P}}\|_{L_\infty(\mathbb{R}^d)}\,,$$

This implies that, for all $(x,y) \in X \times Y$, we have

$$\begin{aligned} L_\tau(y, K * f^*_{L_\tau,\mathrm{P}}(x)) &\leq C_\tau(M + \|K * f^*_{L_\tau,\mathrm{P}}\|_\infty)^2 \\ &\leq 4C_\tau(M + 2^s\|f^*_{L_\tau,\mathrm{P}}\|_{L_\infty(\mathbb{R}^d)})^2 =: B_0\,. \end{aligned}$$

For the second results, we first obtain the following upper bound of the regularization term by using (Eberts and Steinwart, 2013, Theorem 2.3)

$$\|f_0\|_{H_\gamma} = \|K * f^*_{L_\tau,\mathrm{P}}\|_{H_\gamma} \leq (\gamma\sqrt{\pi})^{-\frac{d}{2}}(2^s - 1)\|f^*_{L_\tau,\mathrm{P}}\|_{L_2(\mathbb{R}^d)}.$$

Since P_X has a Lebesgue density $g \in L_2(\mathbb{R}^d)$, by Eberts and Steinwart (2013, Theorem 2.2) the upper bound for L_2-distance between f_0 and $f^*_{L_\tau,\mathrm{P}}$ is

$$\|f_0 - f^*_{L_\tau,\mathrm{P}}\|^2_{L_2(P_X)} = \|K * f^*_{L_\tau,\mathrm{P}} - f^*_{L_\tau,\mathrm{P}}\|^2_{L_2(P_X)} \leq C_{s,2}\,\|g\|_{L_2(\mathbb{R}^d)} c^2 \gamma^{2\alpha}\,, \tag{4.11}$$

where $C_{s,2} :=:= \sum_{i=0}^{\lceil 2s \rceil} \binom{\lceil 2s \rceil}{i} (2d)^{\frac{i}{2}} \prod_{j=1}^{i} (j - \frac{1}{2})^{\frac{1}{2}}$ (see Eberts and Steinwart, 2013, p.27), is the constant only depending on s. Now by using Theorem 3.3 together with (4.11), we obtain

$$\mathcal{R}_{L_\tau,\mathrm{P}}(f_0) - \mathcal{R}^*_{L_\tau,\mathrm{P}} \leq C_\tau \, \|f_0 - f^*_{L_\tau,\mathrm{P}}\|^2_{L_2(P_X)} = C_{\tau,s} \gamma^{2\alpha} \,,$$

where $C_{\tau,s} := c^2 \, C_\tau \, C_{s,2} \, \|g\|_{L_2(\mathbb{R}^d)}$. With these results, we finally obtain

$$\begin{aligned} \inf_{f \in H_\gamma} \lambda \|f\|^2_{H_\gamma} + \mathcal{R}_{L_\tau,\mathrm{P}}(f) - \mathcal{R}^*_{L_\tau,\mathrm{P}} &\leq \lambda \|f_0\|^2_{H_\gamma} + \mathcal{R}_{L_\tau,\mathrm{P}}(f_0) - \mathcal{R}^*_{L_\tau,\mathrm{P}} \,, \\ &\leq C_1 \, \lambda \gamma^{-d} + C_{\tau,s} \, \gamma^{2\alpha} \,, \end{aligned}$$

where $C_1 := (\sqrt{\pi})^{-d} (2^r - 1)^2 \|f^*_{L_\tau,\mathrm{P}}\|^2_{L_2(\mathbb{R}^d)}$. ■

Clearly, the upper bound of the approximation error function in Theorem 4.3 depends on the regularization parameter λ, the kernel width γ, and the smoothness parameter α of the target function $f^*_{L_\tau,\mathrm{P}}$. In order to shrink the right-hand side of (4.10) we need to let $\gamma \to 0$. However, this would let the first term go to infinity unless we simultaneously let $\lambda \to 0$ with a sufficient speed.

4.1.3 Learning Rates for Bounded Regression

In this section, we assume that $Y \subseteq [-M, M]$. With this assumption, we have a concentrated distribution P on $X \times [-M, M]$, that is, $\mathrm{P}(X \times [-M, M]) = 1$. Let us now recall the general oracle inequality given in Theorem 2.20, see also Steinwart and Christmann (2008, Theorem 7.23), which for $\vartheta = 1$ will be used as the basis for establishing oracle inequalities in the case of the L_τ-loss. Note that the oracle inequality in Theorem 2.20 has a constant $K(p)$, and from the proof of (Steinwart and Christmann, 2008, Theorem 7.23)), this constant $K(p)$, for $\vartheta = 1$, can be chosen to be

$$K(p) := \max \left\{ 2700 \cdot 2^{2p} \, C_1^2(p) |L_\tau|^{2p}_{1,M} \, V^{1-p}, 90 \cdot (120)^p \, C_2^{1+p}(p) |L_\tau|^{2p}_{1,M} \, B^{1-p}, 2B \right\}, \qquad (4.12)$$

where the constants $C_1(p)$ and $C_2(p)$ derived from the proof of (Steinwart and Christmann, 2008, Theorem 7.16) are

$$C_1(p) := \frac{2\sqrt{\ln 256} C_p^p}{(\sqrt{2} - 1)(1 - p) 2^{p/2}} \qquad \text{and} \qquad C_2(p) := \left(\frac{8\sqrt{\ln 16} C_p^p}{(\sqrt{2} - 1)(1 - p) 4^p} \right)^{\frac{2}{1+p}}, \qquad (4.13)$$

and by Steinwart and Christmann (2008, Lemma 7.15), we have

$$C_p := \frac{\sqrt{2} - 1}{\sqrt{2} - 2^{\frac{2p-1}{2p}}} \cdot \frac{1 - p}{p}. \qquad (4.14)$$

In addition, $|L_\tau|_{1,M}$, B and V are Lipschitz constant, supremum bound and variance bound, respectively, for the L_τ-loss, see Lemma 3.1 and Lemma 3.4. Although the above given expression for $K(p)$ looks complicated, the following lemma shows that one can easily obtain a nice bound of it for all $p \in (0, \frac{1}{2}]$.

Lemma 4.4. *Let $M > 0$ and $\tau \in (0,1)$, then for all $p \in (0, \frac{1}{2}]$ the constant defined by (4.12) is bounded by*

$$K \leq 2 \cdot 10^9 \, C_\tau^3 \, c_\tau^{-1} \, M^3 \,, \tag{4.15}$$

where $C_\tau > 0$ and $c_\tau > 0$ are defined in Lemma 3.1 and Theorem 3.2.

For the proof of Lemma 4.4, we first need the following lemma

Lemma 4.5. *The function $h : (0, \frac{1}{2}] \to \mathbb{R}$ defined by*

$$h(p) := \left(\frac{\sqrt{2}-1}{\sqrt{2} - 2^{\frac{2p-1}{2p}}} \right)^p \,,$$

is convex. Moreover, we have $\sup_{p \in (0,\frac{1}{2}]} h(p) = 1$.

Proof. Let us consider $t := 2p$. Then it suffices to show that the function $g : (0,1] \to \mathbb{R}$ defined by

$$g(t) := \left(\frac{\sqrt{2}-1}{\sqrt{2} - 2^{1-\frac{1}{t}}} \right)^{\frac{t}{2}} \,,$$

is convex. To solve the latter, we first compute the first and second derivative of $g(t)$ with respect to t, that is:

$$g'(t) = \frac{1}{2} \left(\frac{\sqrt{2}-1}{\sqrt{2} - 2^{1-\frac{1}{t}}} \right)^{\frac{t}{2}} \left(\log \left(\frac{\sqrt{2}-1}{\sqrt{2} - 2^{1-\frac{1}{t}}} \right) + \frac{2^{1-\frac{1}{t}} \log 2}{t\big(\sqrt{2} - 2^{1-\frac{1}{t}}\big)} \right) ,$$

and

$$\begin{aligned} g''(t) = &\left(\frac{\sqrt{2}-1}{\sqrt{2} - 2^{1-\frac{1}{t}}} \right)^{\frac{t}{2}} \left(\frac{1}{2} \log \left(\frac{\sqrt{2}-1}{\sqrt{2} - 2^{1-\frac{1}{t}}} \right) + \frac{2^{1-\frac{1}{t}} \log 2}{2t\big(\sqrt{2} - 2^{1-\frac{1}{t}}\big)} \right)^2 \\ &+ \left(\frac{\sqrt{2}-1}{\sqrt{2} - 2^{1-\frac{1}{t}}} \right)^{\frac{t}{2}} \left(\frac{\big(2^{1-\frac{1}{t}}\big)^2 (\log 2)^2}{2t^3\big(\sqrt{2} - 2^{1-\frac{1}{t}}\big)^2} + \frac{2^{1-\frac{1}{t}} (\log 2)^2}{2t^3\big(\sqrt{2} - 2^{1-\frac{1}{t}}\big)} \right) \end{aligned}$$

Since $t \in (0,1]$, it is not hard to see that all terms in $g''(t)$ are strictly positive. Thus $g''(t) > 0$ and hence $g(t)$ is convex. Furthermore, by convexity of $g(t)$, it is easy to find that

$$\sup_{t \in (0,1]} g(t) = \max\{\lim_{t \to 0} g(t), g(1)\} = 1.$$

■

Proof of Lemma 4.4. In order to bound (4.12), we first bound the constants $C_1(p)$ and $C_2(p)$ defined by (4.13). We start with C_p defined by (4.14) and for $p \in (0, \frac{1}{2}]$ we obtain the following bound

$$C_p^p = \Big(\frac{\sqrt{2}-1}{\sqrt{2}-2^{\frac{2p-1}{2p}}}\Big)^p \Big(\frac{1-p}{p}\Big)^p \le e \max_{p\in(0,\frac{1}{2}]} \Big(\frac{\sqrt{2}-1}{\sqrt{2}-2^{\frac{2p-1}{2p}}}\Big)^p = e\,,$$

where we used $\left(\frac{1-p}{p}\right)^p = \left(\frac{1}{p}-1\right)^p \le e$ for all $p \in (0, \frac{1}{2}]$, and Lemma 4.5. Now the bound for $C_1(p)$ is the following:

$$C_1(p) \le \max_{p\in(0,\frac{1}{2}]} \frac{2\sqrt{\ln 256}\, C_p^p}{(\sqrt{2}-1)(1-p)2^{p/2}} \le \frac{4\,e\,\sqrt{\ln 256}}{\sqrt{2}-1} \max_{p\in(0,\frac{1}{2}]} \frac{1}{2^{p/2}} \le 46\,e\,.$$

Analogously, the bound for the constant $C_2(p)$ is

$$C_2^{1+p}(p) \le \max_{p\in(0,\frac{1}{2}]} \left(\frac{8\sqrt{\ln 16}\, C_p^p}{(\sqrt{2}-1)(1-p)4^p}\right)^2 \le \frac{256\,e^2\ln(16)}{(\sqrt{2}-1)^2} \max_{p\in(0,\frac{1}{2}]} \frac{1}{4^{2p}} \le 1035\,e^2\,.$$

By plugging $C_1(p)$ and $C_2(p)$ together with B and V from Lemma 3.4 and $|L_\tau|_{1,M} = 4\,C_\tau\,M$ from Lemma 3.1 into (4.12), we thus obtain

$$\begin{aligned} K &\le \max\{4\cdot 10^7\, C_\tau^3\, c_\tau^{-1}\, M^3, 2\cdot 10^9\, C_\tau^2 M^3, 8\, C_\tau\, M^2\} \\ &\le 2\cdot 10^9\, C_\tau^3\, c_\tau^{-1}\, M^3\,, \end{aligned}$$

■

Note that in the case of the Gaussian RKHS H_γ, the learning algorithm such as (4.1) constructs a decision function $f_{\mathrm{D},\lambda,\gamma}$. In the following theorem, we establish oracle inequalities for $f_{\mathrm{D},\lambda,\gamma}$ using Theorem 4.3 together with Lemma 3.4 and the entropy number bound (4.6).

Theorem 4.6. *Consider the assumptions of Theorem 4.3 and additionally assume that $Y \subseteq [-M, M]$ for $M \ge 1$. Then, for all $n \ge 1, \varrho \ge 1, \gamma \in (0,1)$ and $\lambda \in (0, e^{-2}]$, the SVM using the RKHS H_γ and the ALS loss function L_τ satisfies*

$$\begin{aligned} &\lambda \|f_{\mathrm{D},\lambda,\gamma}\|_{H_\gamma}^2 + \mathcal{R}_{L_\tau,\mathrm{P}}(\widehat{f}_{\mathrm{D},\lambda,\gamma}) - \mathcal{R}^*_{L_\tau,\mathrm{P}} \\ &\qquad \le C M^3 \Big(\lambda\gamma^{-d} + \gamma^{2\alpha} + (\log\lambda^{-1})^{d+1}\, n^{-1}\gamma^{-d} + n^{-1}\varrho\Big)\,, \end{aligned} \tag{4.16}$$

with probability P^n not less than $1 - 3\,e^{-\varrho}$. Here $C > 0$ is some constant independent of λ, γ, n and ϱ.

Proof of Theorem 4.6. By plugging the results of Theorem 4.3 together with $a = (3K)^{\frac{1}{2p}} \left(\frac{d+1}{ep}\right)^{\frac{d+1}{2p}} \gamma^{-\frac{d}{p}}$ from Theorem 4.2, $B = 4\,C_\tau M^2$ and $V = 16\,c_\tau^{-1}\,C_\tau^2\,M^2$ from Lemma 3.4 into Theorem 2.20, we obtain

$$\begin{aligned}
&\lambda \|f_{\mathrm{D},\lambda,\gamma}\|_{H_\gamma}^2 + \mathcal{R}_{L_\tau,\mathrm{P}}(\widehat{f}_{\mathrm{D},\lambda,\gamma}) - \mathcal{R}^*_{L_\tau,\mathrm{P}} \\
&\leq 9\,C_1 \lambda \gamma^{-d} + 9\,C_{\tau,s}\gamma^{2\alpha} + 3K(p)\,K\Big(\frac{d+1}{e}\Big)^{d+1} \frac{\gamma^{-d}}{p^{d+1}\lambda^p n} \\
&\quad + (3456 M^2\,C_\tau^2\,c_\tau^{-1} + 60(M + 2^s \|f^*_{L_\tau,\mathrm{P}}\|_{L_\infty(\mathbb{R}^d)})^2) \frac{\varrho}{n}\,, \\
&\leq 9\,C_1 \lambda \gamma^{-d} + 9\,C_{\tau,s}\gamma^{2\alpha} + C_d\,K(p)\,\frac{\gamma^{-d}}{p^{d+1}\lambda^p n} + C_2 \frac{\varrho}{n}\,,
\end{aligned} \tag{4.17}$$

where C_1 and $C_{\tau,s}$ are from Theorem 4.3, $K(p)$ is from Lemma 4.4, $C_2 := 3456\,M^2\,C_\tau^2\,c_\tau^{-1} + 60(M + 2^s\|f^*_{L_\tau,\mathrm{P}}\|_{L_\infty(\mathbb{R}^d)})^2$, and $C_d := 3K\Big(\frac{d+1}{e}\Big)^{d+1}$ is a constant only depending on d. Let us assume that $p := \frac{1}{\log \lambda^{-1}}$. Since $\lambda \leq e^{-2}$ and $\lambda^p = e^{-1}$, the result (4.17) becomes

$$\begin{aligned}
&\lambda \|f_{\mathrm{D},\lambda,\gamma}\|_H^2 + \mathcal{R}_{L_\tau,\mathrm{P}}(\widehat{f}_{\mathrm{D},\lambda,\gamma}) - \mathcal{R}^*_{L_\tau,\mathrm{P}} \\
&\leq 9\,C_1 \lambda \gamma^{-d} + 9\,C_{\tau,s}\gamma^{2\alpha} + C_d\,e\,K(p)\,(\log \lambda^{-1})^{d+1}\,\frac{\gamma^{-d}}{n} + C_2 \frac{\varrho}{n}
\end{aligned} \tag{4.18}$$

and by plugging value of $K(p)$ from Lemma 4.4 into (4.18), we finally obtain

$$\begin{aligned}
&\lambda \|f_{D,\lambda,\gamma}\|_H^2 + \mathcal{R}_{L_\tau,\mathrm{P}}(\widehat{f}_{D,\lambda,\gamma}) - \mathcal{R}^*_{L_\tau,\mathrm{P}} \\
&\leq C M^3 \Big(\lambda\gamma^{-d} + \gamma^{2\alpha} + (\log \lambda^{-1})^{d+1}\gamma^{-d} n^{-1} + \varrho\, n^{-1}\Big)\,,
\end{aligned}$$

where C is a constant independent of λ, γ, n and ϱ. ∎

It is well known that there exists a relationship between Sobolev spaces and the scale of Besov spaces, that is, $B^\alpha_{p,u}(\mathbb{R}^d) \hookrightarrow W^\alpha_p(\mathbb{R}^d) \hookrightarrow B^\alpha_{p,v}(\mathbb{R}^d)$, whenever $1 \leq u \leq \min\{p, 2\}$ and $\max\{p, 2\} \leq v \leq \infty$ (see e.g. Edmunds and Triebel, 2008, p.25 and p.44). In particular, for $p = u = v = 2$, we have $W^\alpha_2(\mathbb{R}^d) = B^\alpha_{2,2}(\mathbb{R}^d)$ with equivalent norms. In addition, by Eberts and Steinwart (2013, p.7) we have $W^\alpha_p(\mathbb{R}^d) \subset B^\alpha_{p,q}(\mathrm{P}_X)$. Thus, Theorem 4.6 also holds for decision functions $f^*_{L_\tau,\mathrm{P}} : \mathbb{R}^d \to \mathbb{R}$ with $f^*_{L_\tau,\mathrm{P}} \in L_2(\mathbb{R}^d) \cap L_\infty(\mathbb{R}^d)$ and $f^*_{L_\tau,\mathrm{P}} \in W^\alpha_2(\mathbb{R}^d)$.

In the following corollary we assume some suitable values for λ and γ that depend on data size n, the smoothness parameter α, and the dimension d, and establish learning rates for learning problem (1.7).

Corollary 4.7. *Under the assumptions of Theorem 4.6 and with*

$$\lambda_n = c_1 n^{-1}\,,$$

$$\gamma_n = c_2 n^{-\frac{1}{2\alpha+d}},$$

where $c_1 > 0$ and $c_2 > 0$ are user specified constants, we have, for all $n \geq 1$ and $\varrho \geq 1$,

$$\mathcal{R}_{L_\tau,\mathrm{P}}(\widehat{f}_{\mathrm{D},\lambda,\gamma}) - \mathcal{R}^*_{L_\tau,\mathrm{P}} \leq CM^3\varrho(\log n)^{d+1} n^{-\frac{2\alpha}{2\alpha+d}} \tag{4.19}$$

with probability P^n not less than $1 - 3e^{-\varrho}$. Here $C > 0$ is a constant independent of n and ϱ.

Proof of Corollary 4.7. For all $n \geq 1$, Theorem 4.6 yields

$$\begin{aligned}
&\lambda \|f_{\mathrm{D},\lambda,\gamma}\|^2_{H_\gamma} + \mathcal{R}_{L_\tau,\mathrm{P}}(\widehat{f}_{\mathrm{D},\lambda,\gamma}) - \mathcal{R}^*_{L_\tau,\mathrm{P}} \\
&\leq cM^3(\log \lambda^{-1})^{d+1}\Big(\lambda_n \gamma_n^{-d} + \gamma_n^{2\alpha} + n^{-1}\gamma_n^{-d} + n^{-1}\varrho\Big),
\end{aligned}$$

with probability P^n not less than $1-3e^{-\varrho}$ and a constant $c > 0$. Using the sequences $\lambda_n = c_1 n^{-1}$ and $\gamma_n = c_2 n^{-\frac{1}{2\alpha+d}}$, we obtain

$$\begin{aligned}
&\lambda \|f_{\mathrm{D},\lambda,\gamma}\|^2_{H_\gamma} + \mathcal{R}_{L_\tau,\mathrm{P}}(\widehat{f}_{\mathrm{D},\lambda,\gamma}) - \mathcal{R}^*_{L_\tau,\mathrm{P}} \\
&\leq CM^3(\log n)^{d+1}\Big((c_1 c_2^{-d} + c_2^{2\alpha} + c_2^{-d}) n^{-\frac{2\alpha}{2\alpha+d}} + n^{-1}\varrho\Big) \\
&\leq \tilde{C}M^3\varrho(\log n)^{d+1} n^{-\frac{2\alpha}{2\alpha+d}},
\end{aligned}$$

where the positive constant $\tilde{C} := C(c_1 c_2^{-d} + c_2^{2\alpha} + c_2^{-d} + 1)$ is independent of n and ϱ. ■

4.1.4 Learning Rates using Data Dependent Parameter Selction

We have seen in Corollary 4.7 that learning rates depend on the choice of λ_n and γ_n, where the kernel width γ_n requires knowing α which, in practice, is not available. However, Steinwart and Christmann (2008, Chapter 7.4), Steinwart et al (2009) and Eberts and Steinwart (2013) showed that one can achieve the same learning rates adaptively, i.e. without knowing α. Let us recall (Steinwart and Christmann, 2008, Definition 6.28) that describes a method to select λ and γ in a data-dependent way, which in some sense is a simplification of the cross-validation method.

Definition 4.8. *Let H_γ be a RKHS over X and $\Lambda := (\Lambda_n)$ and $\Gamma := (\Gamma_n)$ be the sequences of finite subsets $\Lambda_n, \Gamma_n \subset (0, 1]$. For a data set $D := ((x_1, y_1), \ldots, (x_n, y_n)) \in (X \times \mathbb{R})^n$, define*

$$\begin{aligned}
D_1 &:= ((x_1, y_1), \ldots, (x_m, y_m)), \\
D_2 &:= ((x_{m+1}, y_{m+1}), \ldots, (x_n, y_n)),
\end{aligned}$$

where $m = \lfloor \frac{n}{2} \rfloor + 1$ *and* $n \geq 4$. *Then use* D_1 *as a training set to compute the SVM decision function*

$$f_{\mathrm{D}_1,\lambda,\gamma} := \arg\min_{f \in H_\gamma} \lambda \|f\|_{H_\gamma}^2 + \mathcal{R}_{L_\tau,\mathrm{D}_1}(f), \qquad (\lambda,\gamma) \in (\Lambda_n, \Gamma_n)\,,$$

and use D_2 *to determine* (λ,γ) *by choosing* $(\lambda_{D_2}, \gamma_{D_2}) \in (\Lambda_n, \Gamma_n)$ *such that*

$$\mathcal{R}_{L_\tau,\mathrm{D}_2}(\widehat{f}_{\mathrm{D}_1,\lambda_{D_2},\gamma_{D_2}}) = \min_{(\lambda,\gamma)\in(\Lambda_n,\Gamma_n)} \mathcal{R}_{L_\tau,\mathrm{D}_2}(\widehat{f}_{\mathrm{D}_1,\lambda,\gamma})\,.$$

Then every learning method that produces the resulting decision functions $\widehat{f}_{\mathrm{D}_1,\lambda_{D_2},\gamma_{D_2}}$ *is called a training validation SVM with respect to* (Λ,Γ).

In the next theorem, we use this training-validation SVM (TV-SVM) approach for suitable candidate sets $\Lambda_n := (\lambda_1,\ldots,\lambda_r)$ and $\Gamma_n := (\gamma_1,\ldots,\gamma_s)$ with $\lambda_r = \gamma_s = 1$, and establish learning rates similar to (4.19).

Theorem 4.9. *With the assumptions of Theorem 4.6, let* $c \geq 2$ *be some constant,* $\Lambda := (\Lambda_n)$ *be a sequence of finite subset* $\Lambda_n \in (0, e^{-2}]$ *such that* $\frac{1}{c} n^{-1} \leq \lambda_i \leq c\, n^{-1}$ *for all* $n \geq 4$ *and* $\Gamma := (\Gamma_n)$ *be sequences of finite subsets* $\Gamma_n \subset (0,1]$ *such that* Γ_n *is an* $n^{-\frac{1}{2\alpha+d}}$*-net of* $(0,1]$. *In addition we assume that the cardinalities* $|\Lambda_n|$ *and* $|\Gamma_n|$ *are polynomially growing in* n. *Then for all* $\varrho \geq 1$, *the TV-SVM produces a* $f_{\mathrm{D}_1,\lambda_{D_2},\gamma_{D_2}}$ *that satisfies*

$$\mathcal{R}_{L_\tau,\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\lambda_{D_2},\gamma_{D_2}}) - \mathcal{R}^*_{L_\tau,\mathrm{P}} \leq C M^3 \varrho (\log n)^{(d+1)}\, n^{-\frac{2\alpha}{2\alpha+d}}$$

with probability P^n *not less than* $1 - 3e^{-\varrho}$, *where* $C > 0$ *is a constant independent of* n *and* ϱ.

In order to prove Theorem 4.9, we first need the following technical lemma.

Lemma 4.10. *Let* $c \geq 3$, $n \geq 3$ *be a constant,* $\Lambda_n \subset (0,1]$ *be a finite set such that there exists a* $\lambda_i \in \Lambda_n$ *with* $\frac{1}{c} n^{-1} \leq \lambda_i \leq c n^{-1}$. *Moreover assume that* $\delta_n \geq 0$ *and* $\Gamma_n \subset (0,1]$ *is a finite* δ_n*-net of* $(0,1]$. *Then for* $d > 0$ *and* $\alpha > 0$ *we have*

$$\inf_{(\lambda,\gamma)\in\Lambda_n\times\Gamma_n} (\lambda\gamma^{-d} + \gamma^{2\alpha} + (\log\lambda^{-1})^{d+1}\gamma^{-d} n^{-1}) \leq c(\log n)^{d+1}\Big(n^{-\frac{2\alpha}{2\alpha+d}} + \delta_n^{2\alpha}\Big)\,,$$

where c *is a constant independent of* $n, \delta_n, \Lambda_n, \Gamma_n$.

Proof. Let us assume that $\Lambda_n = \{\lambda_1,\ldots,\lambda_r\}$ and $\Gamma_n = \{\gamma_1,\ldots,\gamma_s\}$, and $\lambda_{i-1} < \lambda_i$ for all $i = 2,\ldots,r$ and $\gamma_{j-1} < \gamma_j$ for all $j = 2,\ldots,s$. We thus obtain

$$\begin{aligned}
&\inf_{(\lambda,\gamma)\in\Lambda_n\times\Gamma_n} \Big(\lambda\gamma^{-d} + \gamma^{2\alpha} + \frac{(\log\lambda^{-1})^{d+1}}{\gamma^d n}\Big) \\
&\leq \inf_{\gamma\in\Gamma_n} \Big(\lambda_i\gamma^{-d} + \gamma^{2\alpha} + \frac{(\log\lambda_i^{-1})^{d+1}}{\gamma^d n}\Big)
\end{aligned}$$

$$
\begin{aligned}
&\leq \inf_{\gamma\in\Gamma_n}\Big(c\gamma^{-d}n^{-1}+\gamma^{2\alpha}+(\log c+\log n)^{d+1}\gamma^{-d}n^{-1}\Big)\\
&\leq \big(c+(2\log c)^{d+1}\big)\,(\log n)^{d+1}\inf_{\gamma\in\Gamma_n}\Big(\gamma^{-d}n^{-1}+\gamma^{2\alpha}\Big)\\
&\leq \tilde{c}\,(\log n)^{d+1}\inf_{\gamma\in\Gamma_n}\Big(\gamma^{-d}n^{-1}+\gamma^{2\alpha}\Big), \qquad (4.20)
\end{aligned}
$$

where $\tilde{c}:=c+(2\log c)^{d+1}$. It is not hard to see that the function $\gamma\mapsto\gamma^{-d}n^{-1}+\gamma^{2\alpha}$ is optimal at $\gamma^*:=c_1n^{-\frac{1}{2\alpha+d}}$, where $c_1>0$ is a constant which only depends on α and d. Furthermore, with $\gamma_0=0$, we see that $\gamma_j-\gamma_{j-1}\leq 2\delta_n$ for all $j=1,\ldots,s$. In addition, there exits an index $j\in\{1,\ldots,s\}$ such that $\gamma_{j-1}\leq\gamma_n^*\leq\gamma_j$. Consequently, we have $\gamma_n^*\leq\gamma_j\leq\gamma_n^*+2\delta_n$. Using this result in (4.20), we obtain

$$
\begin{aligned}
&\inf_{(\lambda,\gamma)\in\Lambda_n\times\Gamma_n}\Big(\lambda\gamma^{-d}+\gamma^{2\alpha}+\frac{(\log\lambda^{-1})^{d+1}}{\gamma^d n}\Big)\\
&\leq \tilde{c}\,(\log n)^{d+1}\left(\gamma_j^{-d}n^{-1}+\gamma_j^{2\alpha}\right)\\
&\leq \tilde{c}\,(\log n)^{d+1}\left((\gamma_n^*)^{-d}n^{-1}+(\gamma_n^*+2\delta_n)^{2\alpha}\right)\\
&\leq \tilde{c}\,(\log n)^{d+1}\Big((\gamma_n^*)^{-d}n^{-1}+c_\alpha(\gamma_n^*)^{2\alpha}+c_\alpha\delta_n^{2\alpha}\Big)\\
&\leq \tilde{c}_\alpha\,(\log n)^{d+1}\Big((c_1n^{-\frac{1}{2\alpha+d}})^{-d}n^{-1}+(c_1n^{-\frac{1}{2\alpha+d}})^{2\alpha}+\delta_n^{2\alpha}\Big)\\
&\leq c\,(\log n)^{d+1}\Big(n^{-\frac{2\alpha}{2\alpha+d}}+\delta_n^{2\alpha}\Big),
\end{aligned}
$$

where $c:=\tilde{c}_\alpha(c_1^{-d}+c_1^{2\alpha})$ is a constant. ■

Now by using Lemma 4.10, we prove Theorem 4.9 in the following.

Proof of Theorem 4.9. This proof is the repetition of the proof given by Eberts and Steinwart (2013, Theorem 3.6) for the least squares loss. However, for the sake of completeness, we present here in the case of the L_τ-loss. Let us define $m:=\lfloor\frac{n}{2}\rfloor+1\geq\frac{n}{2}$, then for all $(\lambda,\gamma)\in\Lambda_n\times\Gamma_n$, Theorem 4.6 yields

$$
\begin{aligned}
\mathcal{R}_{L_\tau,\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\lambda,\gamma})-\mathcal{R}^*_{L_\tau,\mathrm{P}} &\leq \frac{c_1}{2}\Big(\lambda\gamma^{-d}+\gamma^{2\alpha}+\frac{(\log\lambda^{-1})^{d+1}}{\gamma^d m}+\frac{\varrho}{m}\Big)\\
&\leq c_1\left(\lambda\gamma^{-d}+\gamma^{2\alpha}+\frac{(\log\lambda^{-1})^{d+1}}{\gamma^d n}+\frac{\varrho}{n}\right),
\end{aligned}
$$

with probability P^m not less than $1-3|\Lambda_n\times\Gamma_n|\,e^{-\varrho}$. Now define $n-m\geq\frac{n}{2}-1\geq\frac{n}{4}$ and $\varrho_n:=\varrho+\ln(1+|\Lambda_n\times\Gamma_n|)$, then by using (Steinwart and Christmann, 2008, Theorem 7.2) and Lemma 4.10, we obtain

$$
\begin{aligned}
&\mathcal{R}_{L_\tau,\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\lambda_{D_2},\gamma_{D_2}})-\mathcal{R}^*_{L_\tau,\mathrm{P}}\\
&\leq 6\inf_{(\lambda,\gamma)\in\Lambda_n,\Gamma_n}\Big(\mathcal{R}_{L_\tau,\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\lambda,\gamma})-\mathcal{R}^*_{L_\tau,\mathrm{P}}\Big)+512M^2\,C_\tau^2\,c_\tau^{-1}\frac{\varrho_n}{n-m}
\end{aligned}
$$

$$
\begin{aligned}
&\leq 6c_1 \inf_{(\lambda,\gamma)\in\Lambda_n,\Gamma_n} \Big(\lambda\gamma^{-d} + \gamma^{2\alpha} + \frac{(\log\lambda^{-1})^{d+1}}{\gamma^d n} + \frac{\varrho}{n}\Big) + 2048M^2\, C_\tau^2\, c_\tau^{-1}\frac{\varrho_n}{n} \\
&\leq 6c_1\Big(c(\log n)^{d+1}\big(n^{-\frac{2\alpha}{2\alpha+d}} + \delta_n^{2\alpha}\big)\Big) + 2048M^2\, C_\tau^2\, c_\tau^{-1}\frac{\varrho_n}{n} \\
&\leq \varrho M^2(\log n)^{d+1}(6c_1c + 6cc_1\delta_n^{2\alpha} + 6c_1 + 2048\, C_\tau^2\, c_\tau^{-1}\varrho_n)n^{-\frac{2\alpha}{2\alpha+d}} \\
&\leq c_2 M^3\varrho(\log n)^{d+1}n^{-\frac{2\alpha}{2\alpha+d}}\,,
\end{aligned}
$$

with probability P^n not less than $1 - 3(1 + |\Lambda_n \times \Gamma_n|)e^{-\varrho}$. ■

4.1.5 Learning Rates for Unbounded Noise

So far we have only considered the case of bounded noise with known bounds, that is, $Y \subset [-M, M]$ where $M \geq 1$ is known. In practice, M is usually unknown and in this situation, one can still achieve the same learning rates by simply increasing M slowly. However, more interesting is the case of unbounded noise. In the following we treat this case for a class of distributions for which the tails of the response variable Y has sufficiently fast decay. To be more precise, we assume that for all $\varrho > 1$, there exist constants $c \geq 1$ and $l > 0$ such that

$$
P(\{(x, y) \in X \times Y : |y| \leq c\varrho^l\}) \geq 1 - e^{-\varrho} \tag{4.21}
$$

For instance, if $\mathrm{P}(\,\cdot\,|x) \sim N(\mu(x), 1)$, the assumption (4.21) is satisfied for $l = \frac{1}{2}$, and for the case where $\mathrm{P}(\,\cdot\,|x)$ has the density whose tails decay like $e^{-|t|}$, the assumption (4.21) holds for $l = 1$ (see Eberts and Steinwart, 2013, Example 3.7 and 3.8).

With this additional assumption, we present learning rates for the case of unbounded noise in the following theorem.

Theorem 4.11. *Let $Y \subset \mathbb{R}$ and P be a probability distribution on $\mathbb{R}^d \times Y$ such that $X := \operatorname{supp}\mathrm{P}_X \subset B_{l_2^d}$. Moreover, assume that the τ-expectile $f^*_{L_\tau,\mathrm{P}}$ satisfies $f^*_{L_\tau,\mathrm{P}}(x) \in [-1, 1]$ for P_X-almost all $x \in X$, and both $f^*_{L_\tau,\mathrm{P}} \in L_2(\mathbb{R}^d) \cap L_\infty(\mathbb{R}^d)$ and $f^*_{L_\tau,\mathrm{P}} \in B^\alpha_{2,\infty}(\mathrm{P}_X)$ for some $\alpha \geq 1$. In addition, assume that (4.21) holds for all $\varrho \geq 1$. We define*

$$
\begin{aligned}
\lambda_n &= c_1 n^{-1} \\
\gamma_n &= c_2 n^{-\frac{1}{2\alpha+d}}\,,
\end{aligned}
$$

where $c_1 > 0$ and $c_2 > 0$ are user-specified constants. Moreover, for some fixed $\hat{\varrho} \geq 1$ and $n \geq 3$ we define $\varrho := \hat{\varrho} + \ln n$ and $M_n := 2c\varrho^l$, where c is defined in (4.21). Furthermore, we consider the SVM that clips decision function $f_{\mathrm{D},\lambda_n,\gamma_n}$ at M_n after training. Then there exists a $C > 0$ independent of n, p and $\hat{\varrho}$ such that

$$
\lambda_n\|f_{\mathrm{D},\lambda_n,\gamma_n}\|^2_{H_{\gamma_n}} + \mathcal{R}_{L_\tau,\mathrm{P}}(\widehat{f}_{\mathrm{D},\lambda_n,\gamma_n}) - \mathcal{R}^*_{L_\tau,\mathrm{P}} \leq C\hat{\varrho}^{3l+1}(\log n)^{3l+d+1}n^{-\frac{2\alpha}{2\alpha+d}} \tag{4.22}
$$

holds with probability P^n not less than $1 - 2e^{-\hat{\varrho}}$.

Proof of Theorem 4.11. By (4.21), we obtain

$$\begin{aligned} P^n\Big(\Big\{D \in (X \times Y)^n : \max_{i\in\{1,\ldots,n\}}\{|y_i|\} \le c\varrho^l\Big\}\Big) &\ge 1 - \sum_{i=1}^n P(|y_i| \ge c\varrho^l) \\ &\ge 1 - ne^{-\varrho} \\ &= 1 - e^{-(\varrho - \ln n)}\,. \end{aligned}$$

This implies that

$$P^n\Big(\Big\{D \in (X \times Y)^n : \max_{i\in\{1,\ldots,n\}}\{|y_i|\} \le c(\hat{\varrho} + \ln n)^l\Big\}\Big) \ge 1 - e^{-\hat{\varrho}}\,.$$

This leads us to conclude with probability P^n not less than $1 - e^{-\hat{\varrho}}$ that the SVM for ALS loss with belatedly clipped decision function at M_n is actually a clipped regularized empirical risk minimization (CR-ERM) in the sense of (Steinwart and Christmann, 2008, Definition 7.18). Consequently, (Steinwart and Christmann, 2008, Theorem 7.20) holds for $\hat{Y} := [-M_n, M_n]$ modulo a set of probability P^n not less than $1 - e^{-\hat{\varrho}}$. From Theorem 4.6, we then obtain

$$\begin{aligned} &\lambda\|f_{\mathrm{D},\lambda,\gamma}\|_{H_\gamma}^2 + \mathcal{R}_{L_\tau,\mathrm{P}}(\widehat{f}_{\mathrm{D},\lambda,\gamma}) - \mathcal{R}^*_{L_\tau,\mathrm{P}} \\ &\le CM_n^2\,(\log\lambda^{-1})^{d+1}\Big(\lambda\gamma^{-d} + \gamma^{2\alpha} + n^{-1}\gamma^{-d} + n^{-1}\bar{\varrho}\Big)\,. \end{aligned}$$

with probability P^n not less than $1 - e^{-\bar{\varrho}} - e^{-\hat{\varrho}}$. As in the proof of Corollary (4.7) and by using the inequality $(a+b)^c \le (2ab)^c$, for $a, b \ge 1$ and $c > 0$, we finally obtain

$$\begin{aligned} \lambda\|f_{\mathrm{D},\lambda,\gamma}\|_{H_\gamma}^2 + \mathcal{R}_{L_\tau,\mathrm{P}}(\widehat{f}_{\mathrm{D},\lambda,\gamma}) - \mathcal{R}^*_{L_\tau,\mathrm{P}} &\le C\bar{\varrho}M_n^3(\log n)^{d+1}n^{-\frac{2\alpha}{2\alpha+d}} \\ &= C\bar{\varrho}\left(2c(\hat{\varrho} + \log n)^l\right)^3(\log n)^{d+1}n^{-\frac{2\alpha}{2\alpha+d}} \\ &\le C\bar{\varrho}\,8\,c^3\,(2\hat{\varrho}\,\log n)^{3l}\,(\log n)^{d+1}n^{-\frac{2\alpha}{2\alpha+d}} \\ &\le \hat{C}\bar{\varrho}\hat{\varrho}^{3l}(\log n)^{3l+d+1}n^{-\frac{2\alpha}{2\alpha+d}}\,, \end{aligned}$$

for all $n \ge 3$ with probability P^n not less than $1 - e^{-\bar{\varrho}} - e^{-\hat{\varrho}}$. Choosing $\bar{\varrho} = \hat{\varrho}$ leads to the assertion. ∎

Note that the assumption (4.21) on the tail of the distribution does not influence learning rates achieved in Corollary 4.7. Furthermore, we can also achieve the same rates adaptively using TV-SVM approach considered in Theorem 4.9 provided that we have an upper bound of the unknown parameter l which depends on the distribution P.

4.2 Learning Rates Assuming Generic Kernels

The goal of this section is to establish learning rates of SVM-like algorithm for expectile regression (4.1) in the case of generic kernels. Here we assume that P is a distribution on $X \times Y$ where $X \subset \mathbb{R}^d$ and $Y \subset [-M, M]$ which implies $|\mathrm{P}|_2 = \big(\int_{X\times Y} y^2 \, d\mathrm{P}(x,y) \big)^{1/2} < \infty$, H denotes a seperable RKHS of a bounded measurable kernel k over X, and L_τ is the ALS loss that can be clipped at $[-M, M]$ in the sense of Definition 2.7.

Since $|\mathrm{P}|_2 < \infty$, it is trivial to show that $\mathcal{R}_{L_\tau,\mathrm{P}}(0) < \infty$. By convexity and local Lip-schitz continuity of the L_τ-loss together with (Steinwart and Christmann, 2008, Corollary 5.3), there exists for all $\lambda > 0$ a unique SVM solution $f_{\mathrm{P},\lambda} \in H$. For $f_{\mathrm{P},\lambda}$ we then have

$$\lambda \|f_{\mathrm{P},\lambda}\|_H^2 \le \lambda \|f_{\mathrm{P},\lambda}\|_H^2 + \mathcal{R}_{L_\tau,\mathrm{P}}(f_{\mathrm{P},\lambda}) \le \mathcal{R}_{L_\tau,\mathrm{P}}(0) \,. \tag{4.23}$$

Let us now recall the approximation error function (2.18) and define it in the case of L_τ-loss by

$$\mathcal{A}(\lambda) := \inf_{f\in H} \lambda \|f\|_H^2 + \mathcal{R}_{L_\tau,\mathrm{P}}(f) - \mathcal{R}^*_{L_\tau,\mathrm{P}} \,, \qquad \lambda > 0 \,,$$

Our first goal here is to bound $\mathcal{A}(\lambda)$. For this, we assume that the target function $f^*_{L_\tau,\mathrm{P}} : X \to [-M, M]$ is in a *real interpolation space*. Recall Edmunds and Triebel (2008, Section 1.3.1) that this space can be defined by $\mathbb{K}$-functional, see also Cucker and Zhou (2007, Definition 4.15). Given two Banach spaces $(A, \|\cdot\|_A)$ and $(B, \|\cdot\|_B)$ such that $B \subset A$ and $\mathrm{id} : B \to A$ is continuous, the $\mathbb{K}$-functional $\mathbb{K} : A \times (0,\infty) \to \mathbb{R}$ for $a \in A$ and $t > 0$ is defined by

$$\mathbb{K}(a,t) := \inf_{b\in B} \big(\|a - b\|_A + t\|b\|_B \big) \tag{4.24}$$

Now, for $0 < \beta < 1$ and $1 \le q \le \infty$, the interpolation space $[A,B]_{\beta,q}$ consists of all $a \in A$ such that the norm

$$\|a\|_{\beta,q} := \begin{cases} \big(\int_0^\infty t^{-\beta} \mathbb{K}(a,t)^q \frac{dt}{t} \big) & \text{if} \quad q < \infty \,, \\ \sup_{t>0} (t^{-\beta} \mathbb{K}(a,t)) & \text{if} \quad q = \infty \,, \end{cases}$$

is finite. The limiting cases of $[A,B]_{\beta,q}$ are $[A,B]_{0,\infty} := A$ and $[A,B]_{1,\infty} := B$. Recall Edmunds and Triebel (2008, Section 1.3.3, eq(3)) that for all $0 < \beta < 1$ and $1 \le q \le q' \le \infty$, the interpolation spaces have continuous embedding, that is,

$$[A,B]_{\beta,1} \subset [A,B]_{\beta,q} \subset [A,B]_{\beta,q'} \subset [A,B]_{\beta,\infty} \,.$$

Note that for a fixed $a \in [A,B]_{\beta,q}$, the $\mathbb{K}(a,t)$ is continuous, non-decreasing and bounded by

$$\mathbb{K}(a,t) \le c\, t^\beta \,, \tag{4.25}$$

see Edmunds and Triebel (2008, Section 1.3.3, eq(2)). Clearly $\mathbb{K}(a,t)$ tends to zero as $t \to 0$. Now using Theorem 3.3 and the idea of interpolation spaces, we bound $\mathcal{A}(\lambda)$ in the following.

Lemma 4.12. *Let* P *be a distribution on* $X \times Y$*,* H *be a separable RKHS associated to a bounded measurable kernel* k *over* X *and* $\beta \in (0,1)$*. Then there exists a constant* $\tilde{c} > 0$ *such that*

$$\mathcal{A}(\lambda) \leq \tilde{c}\,\lambda^{\beta} \tag{4.26}$$

holds for all $\lambda > 0$ *if and only if* $f^*_{L_\tau,\mathrm{P}} \in [L_2(P_X), H]_{\beta,\infty}$*.*

Proof of Lemma 4.12. We first assume that $f^*_{L_\tau,P} \in [L_2(P_X), H]_{\beta,\infty}$ for some $\beta \in (0,1)$. By (4.25), there exists a constant $c > 0$ such that

$$\mathbb{K}(f^*_{L_\tau,\mathrm{P}}, \lambda) \leq c\lambda^{\beta}\,,$$

for all $\lambda > 0$. This result together with (4.24) and the left hand side of (3.14) yields

$$\begin{aligned}
\mathcal{A}(\lambda) &= \inf_{f\in H}\{\lambda\|f\|_H^2 + \mathcal{R}_{L_\tau,\mathrm{P}}(f) - \mathcal{R}^*_{L_\tau,\mathrm{P}}\} \\
&\leq \inf_{f\in H}\{\lambda\|f\|_H^2 + \|f - f^*_{L_\tau,\mathrm{P}}\|^2_{L_2(P_X)}\} \\
&\leq \inf_{f\in H}\left(\sqrt{\lambda}\|f\|_H + \|f - f^*_{L_\tau,\mathrm{P}}\|_{L_2(P_X)}\right)^2 \\
&= \mathbb{K}^2(f^*_{L_\tau,\mathrm{P}}, \sqrt{\lambda}) \\
&\leq c^2\lambda^{\beta}\,.
\end{aligned}$$

This shows (4.26). To show the converse implication, we consider the right hand side of (3.14) and obtain

$$\begin{aligned}
\mathbb{K}(f^*_{L_\tau,\mathrm{P}}, \lambda) &= \inf_{f\in H}\{\lambda\|f\|_H + \|f - f^*_{L_\tau,\mathrm{P}}\|_{L_2(P_X)}\} \\
&\leq \sqrt{2}\inf_{f\in H}\left(\lambda^2\|f\|_H^2 + \|f - f^*_{L_\tau,\mathrm{P}}\|^2_{L_2(P_X)}\right)^{1/2} \\
&= \sqrt{2}\left(\inf_{f\in H}\{\lambda^2\|f\|_H^2 + \|f - f^*_{L_\tau,\mathrm{P}}\|^2_{L_2(P_X)}\}\right)^{1/2} \\
&\leq \sqrt{2}\left(\inf_{f\in H}\{\lambda^2\|f\|_H^2 + c_\tau^{-1}\big(\mathcal{R}_{L_\tau,\mathrm{P}}(f) - \mathcal{R}^*_{L_\tau,\mathrm{P}}\big)\}\right)^{1/2} \\
&\leq \sqrt{2}\,c_\tau^{-1/2}\big(\inf_{f\in H}\{\lambda^2\|f\|_H^2 + \mathcal{R}_{L_\tau,\mathrm{P}}(f) - \mathcal{R}^*_{L_\tau,\mathrm{P}}\big)^{1/2} \\
&= \sqrt{2}\,c_\tau^{-1/2}\sqrt{\mathcal{A}(\lambda^2)} \\
&\leq \sqrt{2}\,c_\tau^{-1/2}\tilde{c}^{1/2}\lambda^{\beta}\,,
\end{aligned}$$

where in the last step we used (4.26). This implies that $f^*_{L_\tau,\mathrm{P}} \in [L_2(P_X), H]_{\beta,\infty}$. ∎

The following lemma gives the the supremum bound of L_τ with respect to $f_{\mathrm{P},\lambda}$, that is an important ingredient for establishing oracle inequalities in Theorem 4.14.

Lemma 4.13. *Let $Y \subseteq [-M, M]$, $L_\tau : Y \times \mathbb{R} \to [0, \infty)$ be the ALS loss for fixed $\tau \in (0,1)$ and H be a separable RKHS of a bounded measurable kernel k over X with $\|k\|_\infty \leq 1$. In addition, assume that there exists a constant $C > 0$ and $s \in (0, 1]$ such that*

$$\|f\|_\infty \leq C\|f\|_H^s \|f\|_{L_2(\mathrm{P}_X)}^{1-s}. \tag{4.27}$$

for all $f \in H$. Then for all $\lambda > 0$ and $\tau \in (0,1)$,

$$\|L_\tau \circ f_{\mathrm{P},\lambda}\|_\infty \leq 4C_\tau M^2 + 8C_\tau M^2\, C^2 \left(\frac{\mathcal{A}(\lambda)}{\lambda}\right)^s =: B(\lambda)\,, \tag{4.28}$$

where C_τ is defined in Lemma 3.1.

Proof of Lemma 4.13. For fixed $\tau \in (0,1)$ and $f_{\mathrm{P},\lambda} \in H$, we have

$$\|L_\tau \circ f_{\mathrm{P},\lambda}\|_\infty \leq C_\tau \sup_{y \in [-M,M]} |y - f_{\mathrm{P},\lambda}|^2 \leq 2C_\tau M^2 + 2C_\tau \|f_{\mathrm{P},\lambda}\|_\infty^2\,, \tag{4.29}$$

where $C_\tau = \max\{\tau, 1-\tau\}$. For $\|f_{\mathrm{P},\lambda}\|_\infty^2$, we first compute

$$\begin{aligned}
\|f_{\mathrm{P},\lambda}\|_{L_2(\mathrm{P}_X)}^2 &= \int |f_{\mathrm{P},\lambda}(x)|^2 d\mathrm{P}_X(x) \\
&\leq \int \tau |f_{\mathrm{P},\lambda}(x) - y|^2 + (1-\tau)|f_{\mathrm{P},\lambda}(x) - y|^2 + 2|y|^2 d\mathrm{P}(x,y) \\
&\leq 2\mathcal{R}_{L_\tau,P}(f_{\mathrm{P},\lambda}) + 2M^2 \\
&\leq 4M^2\,,
\end{aligned} \tag{4.30}$$

where we used $\mathcal{R}_{L_\tau,\mathrm{P}}(f_{\mathrm{P},\lambda}) \leq \mathcal{R}_{L_\tau,\mathrm{P}}(0) = M^2$ by (4.23). Now the assumption (4.27) after using (4.30) together with $\lambda \|f_{\mathrm{P},\lambda}\|_H^2 \leq \mathcal{A}(\lambda)$ yields

$$\|f_{\mathrm{P},\lambda}\|_\infty \leq 2^{1-s} M^{1-s} C \left(\frac{\mathcal{A}(\lambda)}{\lambda}\right)^{\frac{s}{2}},$$

and by plugging this result into (4.29), we finally obtain

$$\|L \circ f_{\mathrm{P},\lambda}\|_\infty \leq 4C_\tau M^2 + 8C_\tau M^2\, C^2 \left(\frac{\mathcal{A}(\lambda)}{\lambda}\right)^s.$$

■

Note that the assumption (4.27) is satisfied if and only if the space $[L_2(\mathrm{P}_X), H]_{\beta,1}$ is continuously embedded in $\ell_\infty(X)$, see Bennett and Sharpley (1988, Proposition 2.10). Moreover, it is well known that if $H = W^m(X)$ and P_X is the uniform distribution on X, then

$[L_2(\mathrm{P}_X), W^m(X)]_{\beta,1} = B^{\beta\,m}_{2,1}(X)$, where $B^{\beta\,m}_{2,1}(X)$ is the Besov space and for $\beta = \frac{d}{2m}$ it is continuously embedded in $\ell_\infty(X)$, see Adams and Fournier (2003, p. 230).

We now use the results of Lemma 4.12 and Lemma 4.13 in Theorem 2.20, and formulate the following main oracle inequalities.

Theorem 4.14. *Let* P *be a distribution on* $X \times Y$*,* H *be a separable RKHS of a measurable kernel* k *over* X *with* $\|k\|_\infty < \infty$ *and* $L_\tau : Y \times \mathbb{R} \to [0,\infty)$*,* $\tau \in (0,1)$ *be the ALS loss. In addition, assume that for fixed* $n \geq 1$*, there exist constants* $p \in (0,1)$ *and* $a \geq C_\tau\, 4M^2$ *such that*

$$\mathbb{E}_{D_X \sim P_X^n}\, e_i(\mathrm{id} : H \to L_2(\mathrm{D}_X)) \leq a\, i^{-\frac{1}{2p}}\,, \qquad i \geq 1\,.$$

Moreover, assume that $f^*_{L_\tau,\mathrm{P}} \in [L_2(\mathrm{P}_X), H]_{\beta,\infty}$ *and* (4.27) *hold for some* $\beta \in (0,1)$ *and* $s \in (0,1]$*. Then for all* $\lambda > 0$ *and all* $\varrho > 0$*, the SVM using* H *and* L_τ *satisfies*

$$\begin{aligned}&\lambda \|f_{\mathrm{D},\lambda}\|_H^2 + \mathcal{R}_{L_\tau,\mathrm{P}}(\widehat{f}_{\mathrm{D},\lambda}) - \mathcal{R}^*_{L_\tau,\mathrm{P}} \\ &\leq 9\,\tilde{c}\,\lambda^\beta + K(p)\, a^{2p}\, \lambda^{-p}\, n^{-1} + C_{1,\tau}\, \lambda^{s(\beta-1)}\, n^{-1}\, \varrho + C_{2,\tau}\, n^{-1}\, \varrho \end{aligned} \tag{4.31}$$

with probability P^n *not less than* $1 - 3\,e^{-\varrho}$*, where*

$$\begin{aligned} C_{1,\tau} &:= 120\, M^2\, C^2\, C_\tau \\ C_{2,\tau} &:= (3456\, c_\tau^{-1}\, C_\tau^2 + 60\, C_\tau)\, M^2\,, \end{aligned}$$

and $K(p)$ *is defined by* (4.12)*.*

Proof of Theorem 4.14. Since L_τ is locally Lipschitz continuous loss, there exists a unique function $f_0 = f_{\mathrm{P},\lambda} \in H$ that minimizes $f \mapsto \lambda\|f\|_H + \mathcal{R}_{L_\tau,\mathrm{P}}(f)$. By Lemma 4.12 we then obtain

$$\lambda \|f_{\mathrm{P},\lambda}\|_H + \mathcal{R}_{L_\tau,\mathrm{P}}(f_{\mathrm{P},\lambda}) - \mathcal{R}^*_{L_\tau,\mathrm{P}} = \mathcal{A}(\lambda) \leq \tilde{c}\lambda^\beta\,,$$

for all $\lambda > 0$. In addition, L_τ satisfies the supremum bound (2.14) with $B = 4\, C_\tau\, M^2$ and variance bound (2.15) with $V = 16\, c_\tau^{-1}\, C_\tau^2$, see Lemma 3.4. Furthermore, by Lemma 4.13

$$\|L_\tau \circ f_{\mathrm{P},\lambda}\|_\infty \leq 4C_\tau M^2 + 8C_\tau M^2\, C^2 \left(\frac{\mathcal{A}(\lambda)}{\lambda}\right)^s =: B(\lambda) > B\,.$$

Using all the results in Theorem 2.20, we obtain the assertion. ∎

From Theorem 4.14 we immediately obtain the following learning rates, after assuming appropriate values for λ that depends on β and p.

Corollary 4.15. *Consider $s = p$ in Theorem 4.14 and additionally assume that (4.26) holds for some $\beta \in (0,1)$. Define a sequence of λ_n by*

$$\lambda_n := c\, n^{-\frac{1}{\beta+p}},$$

where $c > 0$ is a user specified constant. Then there exists a constant $K \geq 1$ only depending on M, C, a and p such that for all $\varrho \geq 1$ and $n \geq 1$, the learning method (4.1) satisfies

$$\mathcal{R}_{L_\tau,\mathrm{P}}(\widehat{f}_{\mathrm{D},\lambda_n}) - \mathcal{R}^*_{L_\tau,\mathrm{P}} \leq K\rho n^{-\frac{\beta}{\beta+p}} \tag{4.32}$$

with probability P^n not less than $1 - 3e^{-\varrho\, n^{\frac{p\beta}{\beta+p}}}$.

Proof of Corollary 4.15. Since $s = p$, we rewrite (4.31) for λ_n by

$$\begin{aligned}
&\lambda_n \|f_{\mathrm{D},\lambda_n}\|_H^2 + \mathcal{R}_{L_\tau,\mathrm{P}}(\widehat{f}_{\mathrm{D},\lambda_n}) - \mathcal{R}^*_{L_\tau,\mathrm{P}} \\
&\leq 9\,\tilde{c}\,\lambda_n^\beta + K(p)\, a^{2p}\, \lambda_n^{-p}\, n^{-1} + (C_{1,\tau} + C_{2,\tau})\ \lambda^{p(\beta-1)}\, n^{-1}\, \varrho_n\,.
\end{aligned}$$

By using the sequence $\lambda_n = c\, n^{-\frac{1}{\beta+p}}$ and considering $\varrho_n := \varrho\, n^{\frac{\beta p}{\beta+p}}$ we obtain

$$\begin{aligned}
&\lambda_n \|f_{\mathrm{D},\lambda_n}\|_H^2 + \mathcal{R}_{L_\tau,\mathrm{P}}(\widehat{f}_{\mathrm{D},\lambda_n}) - \mathcal{R}^*_{L_\tau,\mathrm{P}} \\
&\leq 9\,\tilde{c}\, c^\beta\, n^{-\frac{\beta}{\beta+p}} + K(p)\, a^{2p}\, c^{-p}\, n^{-\frac{\beta}{\beta+p}} + (C_{1,\tau} + C_{2,\tau})\ c^{p(\beta-1)}\, \varrho\, n^{-\frac{\beta p-\beta}{\beta+p}+\frac{\beta p}{\beta+p}} \\
&\leq \left(9\,\tilde{c}\, c^\beta + K(p)\, a^{2p}\, c^{-p} + (C_{1,\tau} + C_{2,\tau})\ c^{p(\beta-1)}\right)\ \varrho\, n^{-\frac{\beta}{\beta+p}}\,,
\end{aligned}$$

with probability P^n not less than $1 - 3e^{-\varrho\, n^{\frac{p\beta}{\beta+p}}}$. ∎

Note that the considered choice of λ_n minimizes the right hand side of (4.31) asymptotically, and leads to the fastest learning rates that one can expect. However, we see that learning rates (4.32) depend on p and β, which are unknown in practice. By using TV-SVM for the case of generic kernels, similar to the one presented in Theorem 4.9 for the Gaussian RBF kernels, one can obtain learning rates similar to (4.32) adaptively, that is, without knowing p and β explicitly. Since we have considered $Y \subseteq [-M, M]$, one can restrict the class of distributions similar to the idea presented in Section 4.1.5 and obtain the similar learning rates for unbounded case.

Example 4.16. If we assume that $H = W^m(X)$ for some $m > \frac{d}{2}$, then we have

$$[L_2(\mathrm{P}_X), W^m(X)]_{\beta,\infty} = B^{\beta\, m}_{2,\infty}$$

and we obtain learning rates $n^{-\frac{2\alpha}{2\alpha+d}}$ from (4.32), where $\alpha = \beta\, m \in (d/2, m]$ and we have considered $p = d/2m$. In other words, one can obtain asymptotic optimal learning rates for all regression functions of learning method for expectile regression (4.1) when using Besov space $B^\alpha_{2,\infty}$, $\alpha \in (d/2, m]$ as underlying RKHS H.

4.3 Conclusion

In the first part of this chapter, we established learning rates of SVM-type learning algorithm for expectile regression considering Gaussian RBF kernels. The key ingredients for these learning rates were the bound of approximation error function established in Theorem 4.3, the improved entroy bound of the Gaussian RKHSs derived in Theorem 4.2 and the results that have already been established in Chapter 3. In addition, we derived in this chapter an explicit bound for the constant K for all $p \in (0, \frac{1}{2}]$, whose dependence on p was unknown before. In the second part of this chapter, we considered generic kernels and established learning rates of SVM-type learning algorithm. Here, besides the results of Chapter 3, the bound of approximation error function established in Lemma 4.12 also played a key role.

Let us now compare our results for Gaussian RBF kernels to the oracle inequalities and learning rates established by Eberts and Steinwart (2013) for least squares SVMs. This comparison is justifiable because a) the least squares loss is a special case of L_τ-loss for $\tau = 0.5$, b) the target function $f^*_{L_\tau,\mathrm{P}}$ is assumed to be in the Sobolev or Besov space similar to Eberts and Steinwart (2013), and c) the supremum and the variance bounds for L_τ with $\tau = 0.5$ are the same as the ones used by Eberts and Steinwart (2013). Furthermore, recall that Eberts and Steinwart (2013) used the entropy number bounds (4.5) to control the capacity of the RKHS H_γ which contains a constant $c_{p,d}(X)$ depending on p in an unknown manner. As a result, they obtained a leading constant C in their oracle inequality, (see Eberts and Steinwart, 2013, Theorem 3.1) for which no upper bound can be determined explicitly. We cope with this problem by deriving an improved entropy number bound (4.6) which not only provides the upper bound for $c_{p,d}(X)$ but also helps to determine the value of the constant C in the oracle inequality (4.16) explicitly. As a consequence we improve their learning rates of the form $n^{-\frac{2\alpha}{2\alpha+d}+\xi}$, where $\xi > 0$, by

$$(\log n)^{d+1}\, n^{-\frac{2\alpha}{2\alpha+d}}\,.$$

In other words, the nuisance term n^ξ of learning rates from Eberts and Steinwart (2013) is replaced by the logarithmic term $(\log n)^{d+1}$. Moreover, our learning rates, up to this logarithmic term, are minimax optimal, see e.g. Györfi et al (2002, Chapter 1.7). In addition, our statistical analysis provides learning rates for all asymmetric cases, that is, for $\tau \neq 1/2$, which have not been established in the literature yet, and also were not possible to induce from the work of Eberts and Steinwart (2013).

In the case of generic kernels, for $\tau = 0.5$, we compare our established learning rates to

learning rates obtained by Steinwart et al (2009) for the least squares regression. It turns out that our achieved rates are equal to the ones obtained by Steinwart et al (2009). However, unlike Steinwart et al (2009), we have learning rates not only for the least squares case for which $\tau = 0.5$, but in fact for entire set $\tau \in (0, 1)$.

Chapter 5

An SVM-like Solver for Expectiles Regression

In this chapter, we complement the statsitical analysis given in Chapter 4 for SVM-type learning problem for expectile regression with the empirical results in the case of Gaussian RBF kernels. As we have seen in the previous chapters that expectiles can be described with the help of the asymmetric least square loss function, this link therefore makes it possible to estimate expectiles in a non-parametric framework with a support vector machine (SVM) like approach. For the underlying optimization problem, the main goal of this chapter is to develop an efficient sequential-minimal-optimization-based solver and to provide its convergence analysis. With this aim, we formulate the primal and the dual optimization problem of SVMs for expectile regression in Section 5.1. In Section 5.2, we propose an SMO-type algorithm to update one dual coordinate per iteration along with the stopping criterion and initialization methods. The exact two dimensional optimization problem together with some working set selection strategies is discussed in Section 5.3. Section 5.4 contains the convergence analysis of the solver while some experiments and a discussion on the results can be found in Section 5.5. The detailed experimental results of this chapter can also be found in Appendix. Note that the work of this chapter has already been published in Farooq and Steinwart (2017b). In addition, the source code of the proposed solver for expectile regression (ex-svm) is now a part of *liquidSVM: A Fast and Versatile SVM Package*, and can be downloaded from `http://www.isa.uni-stuttgart.de/software/`

5.1 Primal and Dual Optimization Problem

Let $D := ((x_1, y_1), (x_2, y_2), \cdots (x_n, y_n)) \in (X \times \mathbb{R})^n$ be a training data drawn in an i.i.d. fashion from some unknown distribution P on $X \times Y$, where $X \subset \mathbb{R}^d$ and $Y \subset \mathbb{R}$. In addition, we assume that $f : X \to \mathbb{R}$ is a measurable function. Recall that the empirical L_τ-risk of f is defined by

$$\mathcal{R}_{L_\tau,\mathrm{D}}(f) := \frac{1}{n}\sum_{i=1}^{n} L_\tau(y_i, f(x_i))\,. \tag{5.1}$$

Here, $\mathrm{D} := \frac{1}{n}\sum_{i=1}^{n}\delta_{(x_i,y_i)}$ is the empirical measure associated to the data set D and L_τ is defined by (3.2). Now recall that SVMs construct a predictor $f_{\mathrm{D},\lambda}$ by solving the convex optimization problem of the form

$$f_{\mathrm{D},\lambda} = \arg\min_{f\in H} \lambda\|f\|_H^2 + \mathcal{R}_{L_\tau,\mathrm{D}}(f)\,, \tag{5.2}$$

where $\lambda > 0$ is a regularization parameter and H is a seperable RKHS of a bounded measurable kernel $k : X \times X \to \mathbb{R}$. For input domains $X \subset \mathbb{R}^d$, one often uses SVMs that are equipped with Gaussian RBF kernels. Recall Definition 2.13 that the latter are defined by

$$k_\gamma(x, x') := \exp(-\gamma^{-2}\|x - x'\|_2^2)\,, \qquad x, x' \in \mathbb{R}^d\,, \tag{5.3}$$

where $\gamma > 0$ is called the width parameter. For more properties of Gaussian RBF kernels, we refer to Chapter 2.2. In this chapter, we however, always consider the normalized form of k_γ, that is, for which $k_\gamma(x, x) = 1$ for all $x \in \mathbb{R}^d$. Note that the formulation (5.2) also includes kernels of the form

$$\bar{k}_\gamma(x, x') := \frac{1 + k_\gamma(x, x')}{2}\,,$$

and since the RKHS H_γ of k_γ does not contain constants (see, Steinwart and Christmann, 2008, Corollary 4.44), the RKHS of $\bar{k}_\gamma$ is given by $\bar{H}_\gamma = H_\gamma \oplus \mathbb{R}\mathbb{1}_X$, and elements of $\bar{k}_\gamma$ are of the form $(f, b) \in H_\gamma \times \mathbb{R}$ with $(f, b)(x) = f(x) + b$. In other words, (5.2) includes an SVM formulation with offset b in which the regularized term penalizes b by

$$\lambda\|(f, b)\|_H^2 = \frac{\lambda}{4}\Big(\|f\|_H^2 + b^2\Big)\,.$$

In the following, we always consider regularized problem (5.2), that is without offset form. Note that the SVM-algorithms without offset have been considered in last 15 years because the offset term does in general not promise any theoretical and empirical advantages if a large RKHSs such as Gaussian RKHSs are used, see for instance Vogt (2002), Steinwart (2003), Keerthi et al (2006), Steinwart et al (2011) and references theirin. On the other hand, the

offset term leads to an additional equality constraint in the dual problem (see Cristianini and Shawe-Taylor, 2000, Chapter 6.2), and as a consequence, SMO-type solvers can only update certain pairs of dual variables. In addition, the offset makes it relatively expensive to calculate the duality gap (see Cristianini and Shawe-Taylor, 2000, Chapter 5.3), which may serve as a stopping criterion for these solvers.

To deal with (5.2) algorithmically, we fix a feature space H_0 and a feature map $\Phi : X \to H_0$ of $\mathbb{R}$. Then for all $\mathbf{x} \in X$, one can represent $f \in H$ in terms of $\mathbf{w} \in H_0$ via

$$f(\mathbf{x}) = \langle \mathbf{w}, \phi(\mathbf{x}) \rangle_{H_0} ,$$

see Steinwart and Christmann (2008, Theorem 4.21) for further details. Note that the latter theorem also shows that

$$\|f\|_H = \inf\{\|\mathbf{w}\|_{H_0} : \mathbf{w} \in H_0 \text{ with } f = \langle \mathbf{w}, \phi(\cdot\,) \rangle_{H_0}\} , \tag{5.4}$$

Using (5.1) and (5.4) in the objective function (5.2), we obtain the standard regularized problem for SVMs without offset

$$\arg\min_{\mathbf{w} \in H_0} \lambda \|\mathbf{w}\|_{H_0}^2 + \frac{1}{n} \sum_{i=1}^{n} L_\tau(y_i, f(x_i)) . \tag{5.5}$$

By reformulating (5.5) we obtain the following primal optimization problem

$$\begin{aligned} \arg\min_{\substack{(\mathbf{w}, \xi_+, \xi_-) \\ \mathbf{w} \in H}} \mathcal{P}_C(\mathbf{w}, \xi_+, \xi_-) &:= \frac{1}{2}\|\mathbf{w}\|^2 + C\tau \sum_{i=1}^{n} \xi_{i,+}^2 + C(1-\tau) \sum_{i=1}^{n} \xi_{i,-}^2 , \\ \text{such that} \qquad \xi_{i,+} &\geq y_i - \langle \mathbf{w}, \phi(\mathbf{x}_i) \rangle , \\ \xi_{i,-} &\geq \langle \mathbf{w}, \phi(\mathbf{x}_i) \rangle - y_i , \\ \xi_{i,+} , \xi_{i,-} &\geq 0 , \qquad \forall\, i = 1, \ldots, n , \end{aligned} \tag{5.6}$$

where $C := \frac{1}{2n\lambda} > 0$. Using standard Lagrangian techniques, see e.g. Cristianini and Shawe-Taylor (2000, Chapter 6), one can easily obtain the dual optimization problem

$$\begin{aligned} \arg\max_{(\alpha, \beta)} \mathcal{D}(\alpha, \beta) &:= \langle \alpha - \beta, \mathbf{y} \rangle - \frac{1}{2} \langle \alpha - \beta, K(\alpha - \beta) \rangle - \frac{1}{4C\tau} \langle \alpha, \alpha \rangle - \frac{1}{4C(1-\tau)} \langle \beta, \beta \rangle , \\ \alpha_i &\geq 0, \beta_i \geq 0 , \qquad \forall\, i = 1, \ldots, n . \end{aligned} \tag{5.7}$$

Here $\mathbf{y}$ is the $n \times 1$ vector of labels and K is the $n \times n$ matrix with entries $K_{i,j} := k(x_i, x_j)$, $i, j = 1, \ldots, n$. Note that (5.6) is a convex function as the loss function (3.2) is convex. Analogously, it is not hard to see that the dual optimization problem (5.7) is concave. This ensures the fulfillment of the strong duality assumptions (Cristianini and Shawe-Taylor, 2000, Chapter 5)

and consequently, the primal optimal solution can be obtained from a dual optimal solution using a simple transformation. To be more precise, if (α^*, β^*) is an optimal solution of the dual problem (5.7), then the optimal solution of the corresponding primal problem (5.6) is

$$\begin{aligned} \mathbf{w}^* &:= \sum_{i=1}^{n} (\alpha_i^* - \beta_i^*) \phi(\mathbf{x}_i) \,, \\ \xi_{i,+}^* &:= \max\left\{0, y_i - \langle \mathbf{w}^*, \phi(\mathbf{x}_i) \rangle\right\} , \\ \xi_{i,-}^* &:= \max\left\{0, \langle \mathbf{w}^*, \phi(\mathbf{x}_i) \rangle - y_i\right\} , \end{aligned} \tag{5.8}$$

and by (Steinwart and Christmann, 2008, Lemma 5.1), this primal solution is unique. Furthermore, we obtain for $\mathcal{D}^* := \mathcal{D}(\alpha^*, \beta^*)$ and $\mathcal{P}^* := \mathcal{P}_C^*(\mathbf{w}^*, \xi_+^*, \xi_-^*)$ that $\mathcal{D}^* = \mathcal{P}^*$.

The quadratic nature of (5.7) makes it possible to solve it using quadratic programming (QP) techniques. However, many QP techniques that are used to solve the dual optimization problems, for example, interior point methods (see Wright and Nocedal, 1999; Schölkopf and Smola, 2002) are impractical for large scale problems. Decomposition methods, such as *chunking* (see Vapnik, 2000) have been designed to overcome this problem by breaking the optimization problem into smaller subproblems and solving them iteratively. The limiting case of decomposition methods is the Sequential Minimal Optimization (SMO) methods that optimize two coordinates at each iteration (Platt, 1999) for SVMs with offset and hence, does not require storage of the entire kernel matrix. Section 4 presents this idea in more detail in view of expectile regression without offset. It is also worth noting that SVMs without offset enable us to develop an SMO-type algorithm that updates one coordinate per iteration as a starting point. In the following section, we introduce this algorithm in details.

5.2 Working Set of Size One

Our goal in this section is to develop an SMO-type algorithm that updates a single coordinate in each iteration. For this, we first compute one working set solution from the optimization problem (5.7). Then we establish a rule to select a direction in which the update should be performed as well as a criterion to stop the algorithm. In the end, we present two procedures to initialize the coordinates.

Let us first compute the gradients for α_i and β_i from (5.7) which we will use throughout this section and subsequent sections. By taking partial derivatives of (5.7) w.r.t. α_i and β_i we

obtain

$$\begin{aligned}\nabla\mathcal{D}_{\alpha_i}(\alpha,\beta) &= \langle e_i,\mathbf{y}\rangle - \langle e_i, K(\alpha-\beta)\rangle - \frac{\langle e_i,\alpha\rangle}{2C\tau}\,,\\ \nabla\mathcal{D}_{\beta_i}(\alpha,\beta) &= -\langle e_i,\mathbf{y}\rangle + \langle e_i, K(\alpha-\beta)\rangle - \frac{\langle e_i,\beta\rangle}{2C(1-\tau)}\,.\end{aligned} \tag{5.9}$$

Let us now reformulate (5.7) for one working set solution. For $\alpha,\beta\in\mathbb{R}^n$ and an index $i\in\{1,\ldots,n\}$, we write $\alpha^{\setminus i}:=\alpha-\alpha_i e_i$ and $\beta^{\setminus i}:=\beta-\beta_i e_i$ where e_i is the i-th vector of standard basis of $\mathbb{R}^n$. Then, basic calculations together with assuming $K_{i,i}=1$ for normalized kernels leads to the following dual objective function for 1D-problem

$$\begin{aligned}\mathcal{D}(\alpha^{\setminus i}+\alpha_i e_i,\beta^{\setminus i}+\beta_i e_i) := \mathcal{D}(\alpha^{\setminus i},\beta^{\setminus i}) &+ (\alpha_i-\beta_i)\langle e_i,\mathbf{y}\rangle - \frac{1}{2}(\alpha_i-\beta_i)^2\\ &- (\alpha_i-\beta_i)\langle e_i, K(\alpha^{\setminus i}-\beta^{\setminus i})\rangle - \frac{\alpha_i^2}{4C\tau} - \frac{\beta_i^2}{4C(1-\tau)}\,.\end{aligned} \tag{5.10}$$

Taking partial derivatives of (5.10) w.r.t. α_i and β_i and setting them to zero yield the system of equations

$$\begin{aligned} b_1\alpha_i - \beta_i &= c_i\,,\\ \alpha_i - b_2\beta_i &= c_i\,,\end{aligned} \tag{5.11}$$

where

$$\begin{aligned} b_1 &= \frac{2C\tau+1}{2C\tau}\,,\\ b_2 &= \frac{2C(1-\tau)+1}{2C(1-\tau)}\,,\\ c_i &= \langle e_i,\mathbf{y}\rangle - \langle e_i, K(\alpha^{\setminus i}-\beta^{\setminus i})\rangle = \nabla D_{\alpha_i}(\alpha,\beta) + b_1\langle e_i,\alpha\rangle - \langle e_i,\beta\rangle\,.\end{aligned} \tag{5.12}$$

After solving (5.11), we obtain the global solution

$$\alpha_i^* = \frac{b_2-1}{b_1b_2-1}c_i\,,\qquad \beta_i^* = \frac{1-b_1}{b_1b_2-1}c_i\,. \tag{5.13}$$

Note that $b_1,b_2\in(1,\infty)$ for all $C>0$ and $\tau\in(0,1)$. It is not hard to see from (5.13) that $\alpha_i^*=\beta_i^*=0$ if and only if $c_i=0$. On the other hand, for all $c_i\in\mathbb{R}\setminus\{0\}$, (5.13) leads to the relation

$$\alpha_i^* = -\frac{\tau}{1-\tau}\beta_i^*\,, \tag{5.14}$$

which implies that the global solution (α_i^*,β_i^*) violates the constraints of the dual problem (5.7). In other words, the global maximum of (5.7) does not lie in the set of feasible vectors. The following general theorem describes the way to find the solution in this situation.

Theorem 5.1. *Let $\mathcal{D}:\mathbb{R}^m\to\mathbb{R}$ be a concave and twice continuous differentiable function and $\mathcal{A}\subset\mathbb{R}^m$ be a closed convex set. Assume that there is exactly one $\alpha^*\in\mathbb{R}^m$ with $\mathcal{D}'(\alpha^*)=0$. Then the following statements hold:*

i) For all $\alpha\neq\alpha^*$ we have $\mathcal{D}(\alpha^*)>\mathcal{D}(\alpha)$.

ii) If $\alpha^* \notin \mathcal{A}$, then there exists an $\alpha^\star \in \partial\mathcal{A}$ such that $\mathcal{D}(\alpha^\star) \geq \mathcal{D}(\alpha)$ for all $\alpha \in \mathcal{A}$.

Proof of Theorem 5.1. i) We first show that $\mathcal{D}$ has a global maximum at α^*. To do this, we proceed by contradiction, that is, we assume that there exists an $\alpha \in \mathbb{R}^m$ with

$$\mathcal{D}(\alpha^*) < \mathcal{D}(\alpha)\,. \tag{5.15}$$

By concavity of D, we conclude that for $t \in [0,1]$

$$\mathcal{D}((1-t)\alpha^* + t\alpha) \geq (1-t)\mathcal{D}(\alpha^*) + t\mathcal{D}(\alpha)\,. \tag{5.16}$$

On the other hand, $h := t(\alpha - \alpha^*) \in \mathbb{R}^m$ and Taylor's theorem in the multiple dimensional version yields

$$\begin{aligned}\mathcal{D}((1-t)\alpha^* + t\alpha) &= \mathcal{D}(\alpha^* + h)\,,\\ &= \mathcal{D}(\alpha^*) + \langle \mathcal{D}'(\alpha^*), h\rangle + \frac{1}{2}\langle h, \mathcal{D}''(\alpha^*)h\rangle + O(\| h^2 \|)\,,\\ &= \mathcal{D}(\alpha^*) + \frac{t^2}{2}\langle \alpha - \alpha^*, \mathcal{D}''(\alpha^*)(\alpha - \alpha^*)\rangle + O(t^2)\,.\end{aligned}$$

Using this in (5.16) we obtain

$$\mathcal{D}(\alpha^*) + t\big(\mathcal{D}(\alpha) - \mathcal{D}(\alpha^*)\big) \leq \mathcal{D}(\alpha^*) + \frac{t^2}{2}\langle \alpha - \alpha^*, (\alpha - \alpha^*)\mathcal{D}''(\alpha^*)\rangle + O(t^2)\,, \tag{5.17}$$

and thus

$$c_1 t \leq \frac{c_2}{2}t^2 + O(t^2)\,,$$

where $c_1 := \mathcal{D}(\alpha) - \mathcal{D}(\alpha^*)$ and $c_2 := \langle \alpha - \alpha^*, \mathcal{D}''(\alpha^*)(\alpha - \alpha^*)\rangle$. Furthermore, we have $c_2 \leq 0$ since $\mathcal{D}$ is concave and $c_1 > 0$ by (5.15). For sufficiently small $t > 0$, (5.17) is therefore impossible and hence (5.15) can not be true. Let us now show that $\mathcal{D}$ has no other global maximum. For this, we assume the converse, that is, $\mathcal{D}$ has a global maximum at some $\alpha^{**} \neq \alpha^*$. Then we obtain $\mathcal{D}'(\alpha^{**}) = 0$ by usual calculus, and hence our assumptions are violated. Consequently, $\mathcal{D}$ has its only global maximum at α^*.

ii) If $\alpha^* \notin \mathcal{A}$ then we also have $\alpha^* \notin \mathring{\mathcal{A}}$, where $\mathring{\mathcal{A}}$ denotes the interior of $\mathcal{A}$, and for all $\alpha \in \mathring{\mathcal{A}}$ we thus have $\alpha \neq \alpha^*$. Let us now show that for all $\alpha \in \mathring{\mathcal{A}}$ there exists an $\alpha^\star \in \partial\mathcal{A}$ with

$$\mathcal{D}(\alpha^\star) > \mathcal{D}(\alpha). \tag{5.18}$$

To this end, we fix an $\alpha \in \mathring{\mathcal{A}}$ and consider the function

$$\begin{aligned}\gamma : [0,1] &\to \mathbb{R}^m\\ t &\mapsto (1-t)\alpha^* + t\alpha\,.\end{aligned}$$

Furthermore, we set

$$h := \mathcal{D} \circ \gamma \,.$$

Then it is easy to see that h is concave. Moreover, since $\alpha \neq \alpha^*$, we find $\gamma(t) \neq \alpha^*$ for all $t \in (0,1]$ and thus $h(t) < h(0)$ for all $t \in (0,1]$. By the concavity of h we conclude that h is strictly decreasing. We now show that there exists $t^\star \in (0,1]$ with $\gamma(t^\star) \in \partial\mathcal{A}$. Let us assume the converse, that is, $\Gamma \cap \partial\mathcal{A} = \emptyset$, where $\Gamma := \gamma([0,1])$. Considering the partition $\mathring{\mathcal{A}}$, $\partial\mathcal{A}$, $\mathbb{R}^m \backslash \bar{\mathcal{A}}$, where $\bar{\mathcal{A}}$ denotes the closure of $\mathcal{A}$, we then find by the assumed $\mathcal{A} = \bar{\mathcal{A}}$ and $\Gamma \cap \partial\mathcal{A} = \emptyset$ that

$$\begin{aligned} B_1 &:= \Gamma \cap \mathring{\mathcal{A}} \\ B_2 &:= \Gamma \cap \mathbb{R}^m \backslash \bar{\mathcal{A}} = \Gamma \cap (\mathbb{R}^m \backslash \mathcal{A}), \end{aligned}$$

is a partition of Γ. Since $\alpha \in \mathring{\mathcal{A}}$ and $\alpha^* \notin \mathcal{A}$, we further find $B_1 \neq \emptyset$ and $B_2 \neq \emptyset$. Moreover, since $\mathbb{R}^m \backslash \mathcal{A}$ is open, the sets B_1 and B_2 are relatively open in Γ and Γ. However, the continuous image of a connected set, is connected and thus Γ is connected. This leads to a contradiction, and hence there exists a $t^\star \in [0,1]$ with $\gamma(t^\star) \in \partial\mathcal{A}$. Clearly, we have $t^\star < 1$ since $\alpha \notin \partial\mathcal{A}$. For $\alpha^\star := \gamma(t^\star)$, the already established strict monotonicity of h then shows

$$\mathcal{D}(\alpha^\star) = h(t^\star) > h(1) = \mathcal{D}(\alpha) \,.$$

Consequently we have shown (5.18) and thus

$$\sup_{\alpha \in \partial\mathcal{A}} \mathcal{D}(\alpha) = \sup_{\alpha \in \mathcal{A}} \mathcal{D}(\alpha).$$

In other words, it suffices to show that the supremum over $\partial\mathcal{A}$ is attained at some $\alpha^\star \in \partial\mathcal{A}$. To this end, we first show that $\{\mathcal{D} \geq \rho\}$ is bounded for all $\rho < \mathcal{D}^* := \mathcal{D}(\alpha^*)$. For $\alpha \in S$, where $S \subset \mathbb{R}^m$ denotes the Euclidean unit sphere, we define

$$\begin{aligned} h_\alpha : [0,\infty) &\to \mathbb{R}^m \\ t &\mapsto \mathcal{D}(\alpha^* + t\alpha) \,. \end{aligned}$$

Then h_α is concave and continuously differentiable, and has a global maximum at $t = 0$. Moreover, h_α is strictly decreasing with $\lim_{t\to\infty} h_\alpha(t) = -\infty$. We define

$$t_\alpha := \max\{t \geq 0 : h_\alpha(t) \geq \rho\},$$

where we note that the maximum is indeed attained by the continuity of h_α and $t_\alpha < \infty$. Our next intermediate goal is to show that $\alpha \mapsto t_\alpha$ is continuous. To this end, we fix an $\alpha_0 \in S$, and an $\varepsilon > 0$ with $\sqrt{\varepsilon} < -h'_{\alpha_0}(t_{\alpha_0})$, where we note that $h'_{\alpha_0}(t_{\alpha_0}) < 0$ since h_{α_0} is strictly decreasing

and $\mathcal{D}^* > \rho$. Since $\mathcal{D}$ is continuous differentiable, then there exist a $\delta > 0$ such that for all $\alpha \in S$ with $\|\alpha_0 - \alpha\| \leq \delta$ we have

$$|h'_{\alpha_0}(t_{\alpha_0}) - h'_{\alpha}(t_{\alpha_0})| \leq \varepsilon \,.$$

For $t_\alpha \geq t_{\alpha_0}$, the concavity, or more precisely, the subdifferential inequality of $-h'_\alpha(t_{\alpha_0})$, then gives

$$\begin{aligned} h_\alpha(t_\alpha) &\leq h_\alpha(t_{\alpha_0}) + h'_\alpha(t_{\alpha_0})(t_\alpha - t_{\alpha_0}) \,, \\ &\leq h_{\alpha_0}(t_{\alpha_0}) + \varepsilon + (h'_{\alpha_0}(t_{\alpha_0}) + \varepsilon)(t_\alpha - t_{\alpha_0}) \,, \\ &\leq h_{\alpha_0}(t_{\alpha_0}) + \varepsilon + \frac{1}{2} h'_{\alpha_0}(t_{\alpha_0})(t_\alpha - t_{\alpha_0}) \,. \end{aligned}$$

Now recall that $h_\alpha(t_\alpha) = \rho = h_{\alpha_0}(t_{\alpha_0})$. Thus we obtain

$$0 \leq \varepsilon + \frac{1}{2} h'_{\alpha_0}(t_{\alpha_0})(t_\alpha - t_{\alpha_0}) \,,$$

and since $h'_{\alpha_0}(t_{\alpha_0}) < 0$, we conclude that

$$\frac{-2\varepsilon}{h'_{\alpha_0}(t_{\alpha_0})} \geq t_\alpha - t_{\alpha_0} \,,$$

and thus

$$t_\alpha \leq t_{\alpha_0} + \frac{-2\varepsilon}{h'_{\alpha_0}(t_{\alpha_0})} \leq 2\sqrt{\varepsilon} \,.$$

Since an analogous bound can be established in the case $t \leq t_{\alpha_0}$, we conclude that $\alpha \mapsto t_\alpha$ is continuous. Consequently, there exist an $\alpha_0 \in S$ with $t_{\alpha_0} = \sup_{\alpha \in S} t_\alpha$, and thus $\{\mathcal{D} \geq \rho\}$ is bounded. Now we show that there exist $\alpha^\star \in \mathcal{A}$ with

$$\mathcal{D}^\star := \sup_{\alpha \in \mathcal{A}} \mathcal{D}(\alpha) = \mathcal{D}(\alpha^\star) \,.$$

Clearly there is an $(\alpha_n) \subset \mathcal{A}$ with

$$\mathcal{D}(\alpha_n) \to \mathcal{D}^\star \,,$$

and since $\{\mathcal{D} \geq \rho\}$ is bounded, the sequence α_n is also bounded. Then there is a subsequence α_{n_k} and an $\alpha^\star$ with $\alpha_{n_k} \to \alpha^\star$ and the continuity of $\mathcal{D}$ yields $\mathcal{D}(\alpha_{n_k}) \to \mathcal{D}(\alpha^\star)$. Consequently, we have shown $\mathcal{D}(\alpha^\star) = \mathcal{D}(\alpha)$. Finally $\alpha^\star = \lim \alpha_{n_k} \in \mathcal{A}$ follows from $\mathcal{A} = \bar{\mathcal{A}}$. ■

Recall (5.13) that there exits exactly one pair of (α_i^*, β_i^*) at which the derivative vanishes but (5.14) shows that (α_i^*, β_i^*) is not feasible. In this situation, by Theorem 5.1, the optimal feasible solution can be found on the boundary

$$\{(0, \beta_i) : \beta_i \geq 0\} \cup \{(\alpha_i, 0) : \alpha_i \geq 0\} \,,$$

for all $i \in \{1, \ldots, n\}$. To this end, we split the problem into two cases. In the first case, we plug $\alpha_i = 0$ in (5.10) and then differentiate w.r.t. β_i, which provides

$$\frac{\partial \mathcal{D}(\alpha^{\setminus i}, \beta^{\setminus i} + \beta_i e_i)}{\partial \beta_i} = -\langle e_i, \mathbf{y} \rangle + \langle e_i, K(\alpha^{\setminus i} - \beta^{\setminus i}) \rangle - b_2 \langle e_i, \beta \rangle \,.$$

Setting it to zero gives

$$\alpha_i^+ = 0 \,, \qquad \beta_i^+ = -\frac{c_i}{b_2} \,. \tag{5.19}$$

Analogously, for the second case, plugging $\beta_i = 0$ in (5.10) and differentiating w.r.t. α_i yields

$$\frac{\partial \mathcal{D}(\alpha^{\setminus i} + \alpha_i e_i, \beta^{\setminus i})}{\partial \alpha_i} = \langle e_i, \mathbf{y} \rangle - \langle e_i, K(\alpha^{\setminus i} - \beta^{\setminus i}) \rangle - b_1 \langle e_i, \alpha \rangle \,.$$

Equating it to zero provides

$$\beta_i^+ = 0 \,, \qquad \alpha_i^+ = \frac{c_i}{b_1} \,. \tag{5.20}$$

Since $b_1, b_2 \in (1, \infty)$ are constants for a fixed $\tau \in (0, 1)$, the solution (5.19) and (5.20) solely depend on c_i. If $c_i \neq 0$, then we show in the following theorem that either (5.19) or (5.20) gives the feasible optimal solution.

Theorem 5.2. *For $i \in \{1, \ldots, n\}$, let $c_i \in \mathbb{R}$ and $b_1, b_2 \in (1, \infty)$ be defined by (5.12). Then the following implications holds:*

i) If $c_i < 0$, then (5.19) is the feasible solution.

ii) If $c_i = 0$, then (5.19) and (5.20) are the same feasible solution.

iii) If $c_i > 0$, then (5.20) is the feasible solution.

In particular, exactly one of the two cases (i) and (iii) produces a feasible solution (α_i^+, β_i^+), obtained by

$$\alpha_i^+ = \max\left(0, \frac{c_i}{b_1}\right), \qquad \beta_i^+ = \max\left(0, -\frac{c_i}{b_2}\right).$$

Proof of Theorem 5.2. If $c_i = 0$, it is trivial to prove that (5.20) and (5.20) lead to the same feasible solution. For the case when $c_i \neq 0$, we first assume that $c_i > 0$. Since $b_1, b_2 \in (1, \infty)$, only (5.20) provides a feasible solution because $\beta_i^+ < 0$ in (5.19) for $c_i > 0$. Similarly, if we assume that $c_i < 0$, then $\alpha_i^+ < 0$ in (5.20) while $\beta_i^+ > 0$ in (5.19) which makes it feasible solution. ■

After computing the feasible optimal solution, the next task is to determine the coordinate i in which the update should be performed. Many approaches have been discussed so far for this purpose. A simple approach (see Cristianini and Shawe-Taylor, 2000, p. 132-133) is to update

each coordinate $i = 1, \ldots, n$ iteratively. Another method (see, e.g. Vogt, 2002) is to choose the coordinate that violates the Karush-Kuhn-Tucker (KKT) conditions of optimality the most. The latter approach is implemented in SVM packages, SVM^{light} (Joachims, 1999) and LIBSVM (Chang and Lin, 2011). Another idea, see Steinwart et al (2011), which is followed in this work, is to choose the coordinate i^* whose update achieves the largest improvement for the value of the dual objective function $\mathcal{D}$. In other words, it performs the update in the direction

$$i^* \in \arg \max_{i=1,\ldots,n} \mathcal{D}(\alpha + \delta e_i, \beta + \eta e_i) - \mathcal{D}(\alpha, \beta)\,, \tag{5.21}$$

where $\delta = \alpha_i^+ - \alpha_i$ and $\eta = \beta_i^+ - \beta_i$ denote the difference between the new and the old values of α_i and β_i respectively. Based on this idea, the following lemma establishes a rule to compute the gain in the value of $\mathcal{D}$.

Lemma 5.3. *Let $i \in \{1, \ldots, n\}$, $\alpha, \beta \in \mathbb{R}^n$, and $\delta, \eta \in \mathbb{R}$. Moreover let $b_1, b_2 \in (1, \infty)$ be defined by (5.12), then we have*

$$\begin{aligned} G(\delta, \eta) &:= \mathcal{D}(\alpha + \delta e_i, \beta + \eta e_i) - \mathcal{D}(\alpha, \beta) \\ &= \delta \left(\nabla \mathcal{D}_{\alpha_i}(\alpha, \beta) - \frac{b_1 \delta}{2} \right) + \eta \left(\nabla \mathcal{D}_{\beta_i}(\alpha, \beta) - \frac{b_2 \eta}{2} \right) + \delta\eta\,. \end{aligned} \tag{5.22}$$

Proof of Lemma 5.3. Rewriting the dual objective function (5.7)

$$\begin{aligned} \mathcal{D}(\alpha, \beta) =& \langle \alpha, \mathbf{y} \rangle - \langle \beta, \mathbf{y} \rangle - \frac{1}{2}\langle \alpha, K\alpha \rangle + \langle \alpha, K\beta \rangle - \frac{1}{2}\langle \beta, K\beta \rangle \\ &- \frac{1}{4C\tau}\langle \alpha, \alpha \rangle - \frac{1}{4C(1-\tau)}\langle \beta, \beta \rangle. \end{aligned}$$

and assuming $\delta, \eta \in \mathbb{R}$, the update in $\mathcal{D}$ in the direction i is

$$\mathcal{D}(\alpha + \delta e_i, \beta + \eta e_i) = \mathrm{I} - \mathrm{II} - \mathrm{III}, \tag{5.23}$$

where

$$\begin{aligned} \mathrm{I} &:= \langle \alpha + \delta e_i, \mathbf{y} \rangle - \langle \beta + \eta e_i, \mathbf{y} \rangle\,, \\ &= \langle \alpha - \beta, \mathbf{y} \rangle + \delta \langle e_i, \mathbf{y} \rangle - \eta \langle e_i, \mathbf{y} \rangle\,, \\ \mathrm{II} &:= \frac{1}{2}\langle \alpha + \delta e_i, K(\alpha + \delta e_i) \rangle - \langle \alpha + \delta e_i, K(\beta + \eta e_i) \rangle + \frac{1}{2}\langle \beta + \eta e_i, K(\beta + \eta e_i) \rangle\,, \\ &= \frac{1}{2}\langle \alpha - \beta, K(\alpha - \beta) \rangle + \delta \langle \alpha - \beta, Ke_i \rangle - \eta \langle \alpha - \beta, Ke_i \rangle + \frac{\delta^2}{2} + \frac{\eta^2}{2} - \delta\eta\,, \\ \mathrm{III} &:= \frac{1}{4C\tau}\langle \alpha + \delta e_i, \alpha + \delta e_i \rangle + \frac{1}{4C(1-\tau)}\langle \beta + \eta e_i, \beta + \eta e_i \rangle\,, \\ &= \frac{1}{4C\tau}\langle \alpha, \alpha \rangle + \frac{1}{4C(1-\tau)}\langle \beta, \beta \rangle + \frac{\delta}{2C\tau}\langle \alpha, e_i \rangle + \frac{\eta}{2C(1-\tau)}\langle \beta, e_i \rangle \\ &\quad + \frac{\delta^2}{4C\tau} + \frac{\eta^2}{4C(1-\tau)}\,. \end{aligned}$$

By using b_1, b_2 defined in (5.12) together with I, II and III in (5.23) we obtain

$$\begin{aligned}\mathcal{D}(\alpha+\delta e_i,\beta+\eta e_i)-\mathcal{D}(\alpha,\beta) =&\delta\Big(\langle e_i,\mathbf{y}\rangle-\langle\alpha-\beta,Ke_i\rangle-\frac{\langle e_i,\alpha\rangle}{2C\tau}-\frac{b_1}{2}\delta\Big)\\ &+\eta\Big(\langle e_i,\mathbf{y}\rangle+\langle\alpha-\beta,Ke_i\rangle-\frac{\langle e_i,\beta\rangle}{2C\tau}-\frac{b_2}{2}\eta\Big)+\delta\eta\,,\end{aligned}$$

which yields assertion by using (5.9). ■

In the following, we present Procedure 1 that searches for the best direction using Lemma 5.3 and the feasible solution from Theorem 5.2.

Procedure 1 Calculate $i^* \in \arg\max_{i=1,\dots n}\Big(\mathcal{D}(\alpha+\delta e_i,\beta+\eta e_i)-\mathcal{D}(\alpha,\beta)\Big)$

bestgain $\leftarrow -1$
for $i=1$ to n **do**
  $\delta_i \leftarrow \max\left(0,\frac{c_i}{b_1}\right)-\alpha_i$
  $\eta_i \leftarrow \max\left(0,-\frac{c_i}{b_2}\right)-\beta_i$
  gain $\leftarrow G(\delta_i,\eta_i)$
  if *gain* $>$ *bestgain* **then**
    bestgain $\leftarrow$ *gain*
    $i^* \leftarrow i$
    $\delta_{i^*} \leftarrow \delta_i$
    $\eta_{i^*} \leftarrow \eta_i$
  end if
  return $i^*, \delta_{i^*}, \eta_{i^*}$
end for

5.2.1 Stopping Criteria

It is well-known that solving the problem like (5.7) in an iterative manner requires an appropriate stopping criterion. In the context of SVMs *with offset*, several stopping criteria have been suggested so far. One method is to stop training when the KKT conditions are satisfied up to some predefined tolerance $\epsilon > 0$. Another method is to use the duality gap as a stopping criterion (see, Cristianini and Shawe-Taylor, 2000, p. 109 and 128). The latter method has been used by Steinwart et al (2011) to formulate the duality gap for SVM *without offset*. Following this idea, we define for dual variables $\alpha \in \mathbb{R}_+$ and $\beta \in \mathbb{R}_+$ the set

$$H_{\text{pre}} := \Big\{\sum_{i=1}^{n}(\alpha_i-\beta_i)k(\,\cdot\,,x_i) : n\in\mathbb{N}, \alpha_i,\beta_i\in\mathbb{R}_+, x_i\in X, i=1,\dots,n\Big\}$$

which is dense in H (see Steinwart and Christmann, 2008, Theorem 4.21) and for

$$f_{\alpha,\beta} := \sum_{i=1}^{n} (\alpha_i - \beta_i) k(\,\cdot\,, x_i) \in H_{\text{pre}}\,, \tag{5.24}$$

we have $\|f_{\alpha,\beta}\|_H^2 = \langle \alpha - \beta, K(\alpha - \beta) \rangle$. Note that $f_{\alpha,\beta}$ is the dual representation of $f_{\text{D},\lambda}$ defined by (5.2). Using the dual approximate solution (5.24), we obtain the value of the primal objective function (5.6)

$$\mathcal{P}_C(f_{\alpha,\beta}, \xi_{i,+}, \xi_{i,-}) = \frac{1}{2}\langle \alpha - \beta, K(\alpha - \beta) \rangle + C\tau \sum_{i=1}^{n} \xi_{i,+}^2 + C(1-\tau) \sum_{i=1}^{n} \xi_{i,-}^2\,,$$

and thus the duality gap of $\mathcal{P}_C(f_{\alpha,\beta}, \xi_{i,+}, \xi_{i,-})$ and $\mathcal{D}(\alpha, \beta)$ is obtained by

$$S(\alpha, \beta) := \mathcal{P}_C(f_{\alpha,\beta}, \xi_{i,+}, \xi_{i,-}) - \mathcal{D}(\alpha, \beta)\,. \tag{5.25}$$

For some predefined tolerance $\varepsilon > 0$, duality gap (5.25) tells us to stop the iteration method of solving the problem (5.7) when $S(\alpha, \beta) < \epsilon$. As a result we obtain the ε-approximate solution $f^*_{\alpha,\beta}$ of the true decision function $f_{\text{D},\lambda}$. In order to compute $S(\alpha, \beta)$ efficiently, we split it into

$$\begin{aligned} T(\alpha, \beta) &= \frac{1}{2}\langle (\alpha - \beta), K(\alpha - \beta) \rangle - \mathcal{D}(\alpha, \beta)\,, \\ E(\alpha, \beta) &= \tau \sum_{i=1}^{n} \xi_{i,+}^2 + (1-\tau) \sum_{i=1}^{n} \xi_{i,-}^2\,, \end{aligned} \tag{5.26}$$

such that we have $S(\alpha, \beta) = T(\alpha, \beta) + C \cdot E(\alpha, \beta)$. The value of $T(\alpha, \beta)$ can be obtained at each iteration by updating it in the chosen direction i, such as

$$T(\alpha + \delta e_i, \beta + \eta e_i) = T(\alpha, \beta) - U(\alpha_i, \beta_i, \delta, \eta)\,,$$

where

$$\begin{aligned} U(\alpha_i, \beta_i, \delta, \eta) := \delta &\left(2\nabla \mathcal{D}_{\alpha_i}(\alpha, \beta) + \langle y, e_i \rangle + \frac{\langle \alpha, e_i \rangle}{2C\tau} - \frac{(b_1 + 1)\delta}{2} \right) \\ &+ \eta \left(\nabla \mathcal{D}_{\beta_i}(\alpha, \beta) + \langle y, e_i \rangle + \frac{\langle \beta, e_i \rangle}{2C(1-\tau)} - \frac{(b_2 + 1)\eta}{2} \right) + 2\delta\eta\,. \end{aligned} \tag{5.27}$$

Unlike $T(\alpha, \beta)$, the value of $E(\alpha, \beta)$ can not be updated but needs to be computed from scratch for each iteration. To find an efficient formula we combine (5.6) with (5.24) and obtain

$$\xi_{i,+} = \max\left\{0, \langle y, e_i \rangle - \langle \alpha - \beta, K e_i \rangle\right\} = \max\left\{0, \nabla \mathcal{D}_{\alpha_i}(\alpha, \beta) + \frac{\langle \alpha, e_i \rangle}{2C\tau}\right\},$$

and

$$\xi_{i,-} = \max\left\{0, \langle \alpha - \beta, K e_i \rangle - \langle y, e_i \rangle\right\} = \max\left\{0, -\nabla \mathcal{D}_{\alpha_i}(\alpha, \beta) - \frac{\langle \alpha, e_i \rangle}{2C\tau}\right\}.$$

With these formulas, the computation of $E(\alpha, \beta)$ is an $O(n)$ operations.

Let us now consider a little more involved stopping criterion that looks for the $f^*_{\alpha,\beta} \in H_{\text{pre}}$ for which

$$\lambda \|f^*_{\alpha,\beta}\|^2_H + \mathcal{R}_{L,\mathrm{D}}(\widehat{f}^*_{\alpha,\beta}) \leq \min_{f \in H} \lambda \, \|f\|^2_H + \mathcal{R}_{L,\mathrm{D}}(f) + \epsilon \,, \tag{5.28}$$

see Steinwart and Christmann (2008, Definition 7.18), whereas $\widehat{f}_{\alpha,\beta}$ is clipped at $\pm M \in \mathbb{R}$ in the sense of Definition 2.7. To be more precise, we write the clipped value of $f_{\alpha,\beta} : X \to \mathbb{R}$ at $\pm M$ by

$$\widehat{f}_{\alpha,\beta} = \begin{cases} -M & \text{if } f_{\alpha,\beta} < -M \,, \\ f_{\alpha,\beta} & \text{if } f_{\alpha,\beta} \in [-M, M] \,, \\ -M & \text{if } f_{\alpha,\beta} > -M \,. \end{cases}$$

Based on this idea, the clipped version of (5.24) is

$$\widehat{f}_{\alpha,\beta}(x_i) = \Big[\langle e_i, \mathbf{y} \rangle - \nabla \mathcal{D}_{\alpha_i}(\alpha, \beta) - \frac{\langle \alpha, e_i \rangle}{2C\tau} \Big]^M_{-M} \,, \tag{5.29}$$

which leads to the clipped version of $\xi_{i,+}$ and $\xi_{i,-}$

$$\begin{aligned} \widehat{\xi}_{i,+} &= \max \left\{ 0, \langle \mathbf{y}, e_i \rangle - \Big[\langle e_i, \mathbf{y} \rangle - \nabla \mathcal{D}_{\alpha_i}(\alpha, \beta) - \frac{\langle \alpha, e_i \rangle}{2C\tau} \Big]^M_{-M} \right\} \,, \\ \widehat{\xi}_{i,-} &= \max \left\{ 0, \Big[\langle e_i, \mathbf{y} \rangle - \nabla \mathcal{D}_{\alpha_i}(\alpha, \beta) - \frac{\langle \alpha, e_i \rangle}{2C\tau} \Big]^M_{-M} - \langle \mathbf{y}, e_i \rangle \right\} \,. \end{aligned} \tag{5.30}$$

As a result, we obtain

$$\widehat{E}(\alpha, \beta) := \tau \sum_{i=1}^n \widehat{\xi}^2_{i,+} + (1 - \tau) \sum_{i=1}^n \widehat{\xi}^2_{i,-} \,.$$

Then we see that (5.28) is satisfied if

$$\widehat{S}(\alpha, \beta) := T(\alpha, \beta) + C \cdot \widehat{E}(\alpha, \beta) \leq \frac{\epsilon}{2\lambda} \,. \tag{5.31}$$

Indeed, the use of (5.30) in the stopping criterion (5.31) may provide a substantial decrease in duality gap in each iteration compared to the use of unclipped slack variables, and consequently the learning algorithm may require less number of iterations. Note that one can set $S(\alpha, \beta) \leq \frac{\epsilon}{2\lambda}$ as in (5.31), where ϵ has the same value as in (5.25) (see Steinwart et al, 2007). Furthermore, it is worth noting that, unlike the duality gap stopping criterion for SVM *with offset* given by Cristianini and Shawe-Taylor (2000, p. 109f), both (5.25) and (5.31) are directly computable since they do not require the offset term. In the following, we present an $O(n)$ procedure to update $\nabla \mathcal{D}_\alpha(\alpha, \beta)$, $\nabla \mathcal{D}_\beta(\alpha, \beta)$ and to calculate $S(\alpha, \beta)$. The one for $\widehat{S}(\alpha, \beta)$ is an obvious modification and therefore omitted.

Procedure 2 Update $\nabla\mathcal{D}_{\alpha_i}(\alpha,\beta)$ and $\nabla\mathcal{D}_{\beta_i}(\alpha,\beta)$ in direction i^* and calculate $S(\alpha,\beta)$

$T(\alpha,\beta) \leftarrow T(\alpha,\beta) - U(\alpha_i,\beta_i,\delta,\eta)$

$E(\alpha,\beta) \leftarrow 0$

for $k = 1$ to n **do**

$\nabla\mathcal{D}_{\alpha_k}(\alpha,\beta) \leftarrow \nabla\mathcal{D}_{\alpha_k}(\alpha,\beta) - (\delta-\eta)K_{ik} - \frac{\delta^*}{2C\tau}\delta_{ik}$

$\nabla\mathcal{D}_{\beta_k}(\alpha,\beta) \leftarrow \nabla\mathcal{D}_{\beta_k}(\alpha,\beta) + (\delta-\eta)K_{ik} - \frac{\eta^*}{2C(1-\tau)}\delta_{ik}$

$\xi_{k,+} \leftarrow \max\{0, \nabla\mathcal{D}_{\alpha_k}(\alpha,\beta) + \frac{\alpha_k}{2C\tau}\}$

$\xi_{k,-} \leftarrow \max\{0, -\nabla\mathcal{D}_{\alpha_k}(\alpha,\beta) - \frac{\alpha_k}{2C\tau}\}$

$E(\alpha,\beta) \leftarrow E(\alpha,\beta) + \left(\tau\xi_{k,+}^2 + (1-\tau)\xi_{k,-}^2\right)$

end for

$S(\alpha,\beta) = T(\alpha,\beta) + C \cdot E(\alpha,\beta)$

With all the above calculations, we now summarize the basic idea of the 1D-SVM in Algorithm 1. This algorithm suggests to look repeatedly for the best direction i^* and performs update in that direction until the predefined stopping criterion is satisfied.

Algorithm 1 1D-SVM solver

initialize $\alpha, \beta, \nabla\mathcal{D}_\alpha(\alpha,\beta), \nabla\mathcal{D}_\beta(\alpha,\beta)$ and $T(\alpha,\beta)$

while $S(\alpha,\beta) > \frac{\varepsilon}{2\lambda}$ **do**

$(i^*, \delta_{i^*}, \eta_{i^*}) \leftarrow$ Procedure 1

$\alpha_{i^*} \leftarrow \alpha_{i^*} + \delta_{i^*}$

$\beta_{i^*} \leftarrow \beta_{i^*} + \eta_{i^*}$

use Procedure 2 to update$\nabla\mathcal{D}_\alpha(\alpha,\beta), \nabla\mathcal{D}_\beta(\alpha,\beta)$ in direction i^* by δ_{i^*} and η_{i^*} and calculate $S(\alpha,\beta)$

end while

Clearly, the Algorithm 1 still requires some procedures in order to initialize α and β, and the corresponding gradients. Few procedures for this purpose are given in the following section.

5.2.2 Initialization

Various approaches are available to initialize α and β, and their corresponding gradients. We here briefly describe two approaches, namely, *cold start* and *warm start* that we will use during the implementation of the solver.

W0: Cold Start With Zeros. In this approach, we take $\alpha \leftarrow \mathbf{0}$ and $\beta \leftarrow \mathbf{0}$ to initialize them. The initialization of corresponding gradients and duality gap is given in Procedure 3.

Procedure 3 Initialize by $\alpha \leftarrow 0$, $\beta \leftarrow 0$, compute gradients and dual gap

$E(\alpha, \beta) \leftarrow 0$

for $i = 1$ to n **do**

$\alpha_i \leftarrow 0$

$\beta_i \leftarrow 0$

$\nabla \mathcal{D}_{\alpha_i}(\alpha, \beta) \leftarrow \nabla \mathcal{D}_{\alpha_i}(\alpha^*, \beta^*) + y_i$

$\nabla \mathcal{D}_{\beta_i}(\alpha, \beta) \leftarrow \nabla \mathcal{D}_{\beta_i}(\alpha^*, \beta^*) - y_i$

$\xi_{i,+} \leftarrow \max(0, y_i)$

$\xi_{i,-} \leftarrow \max(0, -y_i)$

$E(\alpha, \beta) \leftarrow E(\alpha, \beta) + \left(\tau \xi_{i,+}^2 + (1-\tau)\xi_{i,-}^2\right)$

end for

$T(\alpha, \beta) \leftarrow 0$

$S(\alpha, \beta) \leftarrow C^{\text{new}} \cdot E(\alpha, \beta)$

W1: Warm Start by Recycling Old Solution. Typically, the hyper-parameter λ is chosen by a search over a grid $\Lambda = \{\lambda_1, \ldots, \lambda_m\}$ of candidates values. If these values are ordered in the form $\lambda_1 > \ldots > \lambda_m$ and the SVM is trained in this order, then the resulting $C^{(1)}, \ldots, C^{(m)}$ satisfy the property that $C^{(j)} < C^{(j+1)}$ for all $j = 1, \ldots, m-1$. For $C^{(1)}$ we initialize the solver with the above mentioned cold start and for $j \geq 2$, we initialize it with a warm start, that is, by taking $\alpha \leftarrow \alpha^*$ and $\beta \leftarrow \beta^*$ where α^*, β^* is the approximate solution obtained by training

Procedure 4 Initialize by $\alpha \leftarrow \alpha^*$, $\beta \leftarrow \beta^*$, compute gradients and dual gap

$E(\alpha, \beta) \leftarrow 0$

for $i = 1$ to n **do**

$\alpha_i \leftarrow \alpha_i^*$

$\beta_i \leftarrow \beta_i^*$

$\nabla \mathcal{D}_{\alpha_i}(\alpha, \beta) \leftarrow \nabla \mathcal{D}_{\alpha_i}(\alpha^*, \beta^*) + \frac{\alpha_i^*}{2\tau}\left(\frac{1}{C^{\text{old}}} - \frac{1}{C^{\text{new}}}\right)$

$\nabla \mathcal{D}_{\beta_i}(\alpha, \beta) \leftarrow \nabla \mathcal{D}_{\beta_i}(\alpha^*, \beta^*) + \frac{\beta_i^*}{2(1-\tau)}\left(\frac{1}{C^{\text{old}}} - \frac{1}{C^{\text{new}}}\right)$

$\xi_{i,+} \leftarrow \max\left(0, \nabla \mathcal{D}_{\alpha_i}(\alpha, \beta) + \frac{\alpha_i}{2\tau C^{\text{new}}}\right)$

$\xi_{i,-} \leftarrow \max\left(0, -\nabla \mathcal{D}_{\alpha_i}(\alpha, \beta) - \frac{\alpha_i}{2\tau C^{\text{new}}}\right)$

$E(\alpha, \beta) \leftarrow E(\alpha, \beta) + \left(\tau \xi_{i,+}^2 + (1-\tau)\xi_{i,-}^2\right)$

end for

$T(\alpha, \beta) \leftarrow T(\alpha, \beta) - \frac{1}{4}\left(\frac{1}{C^{\text{old}}} - \frac{1}{C^{\text{new}}}\right) \sum_{i=1}^n \left(\frac{\alpha_i^2}{\tau} + \frac{\beta_i^2}{1-\tau}\right)$

$S(\alpha, \beta) \leftarrow T(\alpha, \beta) + C^{\text{new}} \cdot E(\alpha, \beta)$

with $C^{old} = C^{j-1}$. Obviously, in this case, we can also recycle parts of $\nabla\mathcal{D}_\alpha(\alpha,\beta)$, $\nabla\mathcal{D}_\beta(\alpha,\beta)$ and $S(\alpha,\beta)$. This is described in in Procedure 4.

5.3 Working Set of Size Two

In this section, we extend the idea presented in the previous section and develop an algorithm to perform an update for *two* coordinates per iteration. For this, we first solve the 2D-problem exactly in Section 5.3. Then we formulate a low cost working set selection strategy in Section 5.3.2 using both the 1D-solution and the 2D-solution. In the end of this section, we establish a stopping criterion for the 2D-problem.

5.3.1 Exact Solution of Two Dimensional Problem

Let us fix two coordinates $i, j \in \{1,\dots,n\}$ with $i \neq j$. We further assume that e_i and e_j are the i-th and j-th vectors of standard basis of $\mathbb{R}^n$, and write $\alpha^{\setminus i,j} := \alpha - \alpha_i e_i - \alpha_j e_j$ and $\beta^{\setminus i,j} := \beta - \beta_i e_i - \beta_j e_j$. By this and using $K_{ii} = K_{jj} = 1$ for normalized kernels, the dual objective function for 2D-problem is

$$\begin{aligned}\tilde{\mathcal{D}} &:= \mathcal{D}(\alpha^{\setminus i,j} + \alpha_i e_i + \alpha_j e_j, \beta^{\setminus i,j} + \beta_i e_i + \beta_j e_j) \\ &= \mathcal{D}(\alpha^{\setminus i,j}, \beta^{\setminus i,j}) + \mathcal{D}(\alpha_i,\beta_i) + \mathcal{D}(\alpha_j,\beta_j) - (\alpha_i-\beta_i)(\alpha_j-\beta_j)K_{ij}\,,\end{aligned} \tag{5.32}$$

where

$$\begin{aligned}\mathcal{D}(\alpha_i,\beta_i) &:= (\alpha_i-\beta_i)\langle e_i,\mathbf{y}\rangle - (\alpha_i-\beta_i)\langle e_i, K(\alpha^{\setminus i,j}-\beta^{\setminus i,j})\rangle - \frac{1}{2}(\alpha_i-\beta_i)^2 \\ &\quad - \frac{1}{4C\tau(1-\tau)}((1-\tau)\alpha_i^2 + \tau\beta_i^2)\,, \\ \mathcal{D}(\alpha_j,\beta_j) &:= (\alpha_j-\beta_j)\langle e_j,\mathbf{y}\rangle - (\alpha_j-\beta_j)\langle e_j, K(\alpha^{\setminus i,j}-\beta^{\setminus i,j})\rangle - \frac{1}{2}(\alpha_j-\beta_j)^2 \\ &\quad - \frac{1}{4C\tau(1-\tau)}((1-\tau)\alpha_j^2 + \tau\beta_j^2)\,.\end{aligned}$$

Taking partial derivatives of (5.32) w.r.t. $\alpha_i, \alpha_j, \beta_i$ and β_j, we obtain the gradients

$$\begin{aligned}\nabla\tilde{\mathcal{D}}_{\alpha_i} &= \langle e_i,\mathbf{y}\rangle - \langle e_i, K(\alpha^{\setminus i,j}-\beta^{\setminus i,j})\rangle - b_1\alpha_i + \beta_i - (\alpha_j-\beta_j)K_{i,j}\,, \\ \nabla\tilde{\mathcal{D}}_{\beta_i} &= -\langle e_i,\mathbf{y}\rangle + \langle e_i, K(\alpha^{\setminus i,j}-\beta^{\setminus i,j})\rangle + \alpha_i - b_2\beta_i + (\alpha_j-\beta_j)K_{i,j}\,, \\ \nabla\tilde{\mathcal{D}}_{\alpha_j} &= \langle e_j,\mathbf{y}\rangle - \langle e_j, K(\alpha^{\setminus i,j}-\beta^{\setminus i,j})\rangle - b_1\alpha_j + \beta_j - (\alpha_i-\beta_i)K_{i,j}\,, \\ \nabla\tilde{\mathcal{D}}_{\beta_j} &= -\langle e_j,\mathbf{y}\rangle + \langle e_j, K(\alpha^{\setminus i,j}-\beta^{\setminus i,j})\rangle + \alpha_j - b_2\beta_j + (\alpha_i-\beta_i)K_{i,j}\,,\end{aligned} \tag{5.33}$$

where b_1, b_2 are defined in (5.12). By setting partial derivatives (5.33) to zero, we obtain the following system of equations

$$\begin{aligned} b_1\alpha_i - \beta_i + k\alpha_j - k\beta_j &= c_i\,, \\ \alpha_i - b_2\beta_i + k\alpha_j - k\beta_j &= c_i\,, \\ k\alpha_i - k\beta_i + b_1\alpha_j - \beta_j &= c_j\,, \\ k\alpha_i - k\beta_i + \alpha_j - b_2\beta_j &= c_j\,, \end{aligned} \tag{5.34}$$

where

$$\begin{aligned} k &:= K_{ij}\,, \\ c_i &:= \langle e_i, \mathbf{y}\rangle - \langle e_i, K(\alpha^{\setminus i,j} - \beta^{\setminus i,j})\rangle\,, \\ &= \nabla\mathcal{D}_{\alpha_i}(\alpha,\beta) + b_1\langle\alpha, e_i\rangle - \langle\beta, e_i\rangle + \langle\alpha-\beta, e_j\rangle k\,, \\ c_j &:= \langle e_j, \mathbf{y}\rangle - \langle e_j, K(\alpha^{\setminus i,j} - \beta^{\setminus i,j})\rangle \\ &= \nabla\mathcal{D}_{\alpha_j}(\alpha,\beta) + b_1\langle\alpha, e_j\rangle - \langle\beta, e_j\rangle + \langle\alpha-\beta, e_i\rangle k\,. \end{aligned}$$

Let $\alpha_i^*, \alpha_j^*, \beta_i^*$ and β_j^* be the solution of (5.34). Then solving (5.34) by matrix operations leads to the following global solution

$$\begin{aligned} |M|\alpha_i^* &= (b_2-1)(b_1b_2-1)c_i + (1-b_2)(b_1+b_2-2)kc_j\,, \\ |M|\beta_i^* &= (1-b_1)(b_1b_2-1)c_i + (b_1-1)(b_1+b_2-2)kc_j\,, \\ |M|\alpha_j^* &= (b_2-1)(b_1b_2-1)c_j + (1-b_2)(b_1+b_2-2)kc_i\,, \\ |M|\beta_j^* &= (1-b_1)(b_1b_2-1)c_j + (b_1-1)(b_1+b_2-2)kc_i\,. \end{aligned} \tag{5.35}$$

Here

$$|M| := b_1^2(b_2^2-k^2) - 2b_1(b_2k^2 + b2 - 2k^2) - (b_2-2)^2k^2 + 1\,,$$

is always positive. This is shown in the following lemma

Lemma 5.4. *For $b_1, b_2 \in (1,\infty)$ and $|k| \le 1$, we have $|M| > 0$.*

Proof of Lemma 5.4. After simplification, we write

$$|M| = (b_1b_2-1)^2 - (b_1+b_2-2)^2k^2\,,$$

and by plugging $b_1 = \frac{2C\tau+1}{2C\tau}$ and $b_2 = \frac{2C(1-\tau)+1}{2C(1-\tau)}$ we obtain

$$|M| = \left(\frac{1}{2C\tau(1-\tau)}\right)^2\left(\left(\frac{2C+1}{2C}\right)^2 - k^2\right).$$

Since $C > 0$, we have $\frac{2C+1}{2C} > 1$. From latter together with $|k| \le 1$ and $\tau \in (0,1)$, we obtain the assertion. ■

Note that, in the case of $c_i = c_j = 0$, we have $\alpha_i^* = \beta_i^* = \alpha_j^* = \beta_j^* = 0$. On the other hand, if $c_i \neq 0$ and/or $c_j \neq 0$, then (5.35) together with Lemma 5.4, after some calculations, leads to the following equations

$$\alpha_i^* = -\frac{\tau}{1-\tau}\beta_i^*\,, \qquad \alpha_j^* = -\frac{\tau}{1-\tau}\beta_j^*\,.$$

Since $\tau \in (0,1)$, the global solution (5.35) thus violates the constraints of (5.7) if and only if $c_i \neq 0$ and/or $c_j \neq 0$, that is, the solution is not feasible. To obtain a feasible solution, we know by Theorem 5.1 that we need to look at the boundaries of the feasible region. In our case, this means that we need to set some of the dual variables to zero. Note that this is an extension of the idea that is presented in the 1D-problem. Let us begin by setting one dual variable to zero, say $\alpha_i = 0$. The resultant expressions of gradients w.r.t. the remaining variables can be deduced from the last three expressions of (5.33) after setting $\alpha_i = 0$. By setting these gradients to zero, we obtain the following system of equations

$$\begin{aligned} -b_2\beta_i + k\alpha_j - k\beta_j &= c_i\,,\\ -k\beta_i + b_1\alpha_j - \beta_j &= c_j\,,\\ -k\beta_i + \alpha_j - b_2\beta_j &= c_j\,, \end{aligned} \tag{5.36}$$

where k, c_i, c_j, b_1 and b_2 are the same as in (5.34). Let us write α_j^+, β_i^+ and β_j^+ be the solution of (5.36). Then, by subtracting the last two equations of (5.36), we obtain

$$\alpha_j^+ = -\frac{\tau}{1-\tau}\beta_j^+\,, \tag{5.37}$$

and hence this solution is again not feasible. In a similar way, setting $\beta_i = 0$ provides the following system of equations

$$\begin{aligned} b_1\alpha_i + k\alpha_j - k\beta_j &= c_i\,,\\ k\alpha_i + b_1\alpha_j - \beta_j &= c_j\,,\\ k\alpha_i + \alpha_j - b_2\beta_j &= c_j\,, \end{aligned}$$

which again leads to (5.37) and thus the same conclusion. The remaining two cases where $\alpha_j = 0$ and $\beta_j = 0$ can be treated analogously. Let us now consider the scenario where two variables are set to be zero. For this, we split the problem into six subcases. First we consider the subcase where we set $\alpha_i = 0$ and $\beta_i = 0$ in (5.32). Taking derivatives w.r.t. α_j and β_j, we have

$$\begin{aligned} \nabla\mathcal{D}_{\alpha_j}(\alpha^{\setminus i}, \beta^{\setminus i}) &= \langle e_j, \mathbf{y}\rangle - \langle e_j, K(\alpha^{\setminus i,j} - \beta^{\setminus i,j})\rangle + \beta_j - b_1\alpha_j\,,\\ \nabla\mathcal{D}_{\beta_j}(\alpha^{\setminus i}, \beta^{\setminus i}) &= -\langle e_j, \mathbf{y}\rangle + \langle e_j, K(\alpha^{\setminus i,j} - \beta^{\setminus i,j})\rangle + \alpha_j - b_2\beta_j\,. \end{aligned} \tag{5.38}$$

Setting (5.38) to zero, we obtain the system of equations

$$\begin{aligned} b_1\alpha_j - \beta_j &= c_j\,, \\ \alpha_j - b_2\beta_j &= c_j\,. \end{aligned} \tag{5.39}$$

Let α_j^+ and β_j^+ be the solution of (5.39). Then subtracting equations of (5.39) leads to

$$\alpha_j^+ = -\frac{\tau}{1-\tau}\beta_j^+, \tag{5.40}$$

which shows that the solution is not feasible. Analogously, the second subcase where $\alpha_j = 0$ and $\beta_j = 0$ leads to the same conclusion. In the third subcase, we set $\alpha_i = 0$ and $\alpha_j = 0$ in (5.32) and differentiate w.r.t. β_i and β_j which gives

$$\begin{aligned} \nabla\mathcal{D}_{\beta_i}(\alpha^{\setminus i,j},\beta) &= -\langle e_i,\mathbf{y}\rangle + \langle e_i, K(\alpha^{\setminus i,j} - \beta^{\setminus i,j})\rangle - \beta_j K_{ij} - b_2\beta_i\,, \\ \nabla\mathcal{D}_{\beta_j}(\alpha^{\setminus i,j},\beta) &= -\langle e_j,\mathbf{y}\rangle + \langle e_j, K(\alpha^{\setminus i,j} - \beta^{\setminus i,j})\rangle - \beta_i K_{ij} - b_2\beta_j\,. \end{aligned} \tag{5.41}$$

Setting (5.41) to zero, we obtain a system of equations which, after some calculations, provides the solution

$$\alpha_i^+ = 0\,, \quad \alpha_j^+ = 0\,, \quad \beta_i^+ = |B_1|^{-1}(kc_j - b_2c_i)\,, \quad \beta_j^+ = |B_1|^{-1}(kc_i - b_2c_j), \tag{5.42}$$

where $|B_1| := b_2^2 - k^2 > 0$. Considering the forth subcase, we set $\beta_i = 0$ and $\beta_j = 0$. Analogous to third subcase, the gradients are

$$\begin{aligned} \nabla\mathcal{D}_{\alpha_i}(\alpha,\beta^{\setminus i,j}) &= \langle e_i,\mathbf{y}\rangle - \langle e_i, K(\alpha^{\setminus i,j} - \beta^{\setminus i,j})\rangle - \alpha_j K_{ij} - b_1\alpha_i\,, \\ \nabla\mathcal{D}_{\alpha_j}(\alpha,\beta^{\setminus i,j}) &= \langle e_j,\mathbf{y}\rangle - \langle e_j, K(\alpha^{\setminus i,j} - \beta^{\setminus i,j})\rangle - \alpha_i K_{ij} - b_1\alpha_j\,, \end{aligned}$$

which leads to the solution

$$\beta_i^+ = 0\,, \quad \beta_j^+ = 0\,, \quad \alpha_i^+ = |B_2|^{-1}(b_1c_i - kc_j)\,, \quad \alpha_j^+ = |B_2|^{-1}(b_1c_j - kc_i)\,, \tag{5.43}$$

where $|B_2| := b_1^2 - k^2 > 0$. For fifth subcase, we set $\alpha_i = 0$ and $\beta_j = 0$ and obtain the following solution

$$\alpha_i^+ = 0\,, \quad \beta_j^+ = 0\,, \quad \beta_i^+ = |B_3|^{-1}(b_1c_i - kc_j)\,, \quad \alpha_j^+ = |B_3|^{-1}(kc_i - b_2c_j)\,, \tag{5.44}$$

where $|B_3| := k^2 - b_1b_2 < 0$. Finally, for the last subcase where $\alpha_j = 0$ and $\beta_i = 0$, the solution can be obtained by interchanging i with j in the solution of fifth subcase, which is

$$\beta_i^+ = 0\,, \quad \alpha_j^+ = 0\,, \quad \alpha_i^+ = |B_3|^{-1}(kc_j - b_2c_i)\,, \quad \beta_j^+ = |B_3|^{-1}(b_1c_j - kc_i)\,. \tag{5.45}$$

It is interesting to note that the solutions (5.42), (5.43), (5.44) and (5.45) have the following common expressions

$$T_1 := kc_j - b_2c_i\,, \tag{5.46}$$

$$T_2 := kc_i - b_2 c_j \,, \tag{5.47}$$

$$T_3 := b_1 c_i - k c_j \,, \tag{5.48}$$

$$T_4 := b_1 c_j - k c_i \,. \tag{5.49}$$

The following lemma investigates the behavior of the above four expressions.

Lemma 5.5. *Assume that $c_i \neq 0$ or $c_j \neq 0$. Then the following implications hold:*

i) If $T_1 \geq 0$ and $T_2 \geq 0$ then we have $c_i < 0$ and $c_j < 0$.

ii) If $T_3 \geq 0$ and $T_4 \geq 0$ then we have $c_i > 0$ and $c_j > 0$.

In particular, the expressions T_1, T_2, T_3 and T_4 are not simultaneously positive or negative.

Proof of Lemma 5.5. i) Since $T_1 \geq 0$ and $T_2 \geq 0$, we have $\frac{b_2}{k} c_i \leq c_j \leq \frac{k}{b_2} c_i$. Since we assumed that $c_i \neq 0$ or $c_j \neq 0$, we conclude from the latter and $b_2, k \geq 0$ that we actually have $c_i \neq 0$ and $c_j \neq 0$. Moreover, $b_2 > 1$ and $k \leq 1$ shows that $c_i < 0$ and $c_j < 0$.

ii) Since $T_3 \geq 0$ and $T_4 \geq 0$, we have $\frac{k}{b_1} c_i \leq c_j \leq \frac{b_1}{k} c_i$. Since we assumed that $c_i \neq 0$ or $c_j \neq 0$, we conclude from the latter and $b_1, k \geq 0$ that we actually have $c_i \neq 0$ and $c_j \neq 0$. Moreover, $b_1 > 1$ and $k \leq 1$ shows that $c_i > 0$ and $c_j > 0$.

Finally, this leads to conclude that T_1, T_2, T_3 and T_4 are not simultaneously positive. By similar arguments, it can be shown that these expressions are not simultaneously negative. ■

The Lemma 5.5 leads to the following theorem which shows that only one case from (5.42), (5.43), (5.44) and (5.45) provides the feasible optimal solution.

Theorem 5.6. *Assume that $c_i \neq 0$ or $c_j \neq 0$, then exactly one of the four cases (5.42), (5.43), (5.44) and (5.45) produces a feasible solution. Moreover, the following implications hold:*

i) If $T_1 \geq 0$ and $T_2 \geq 0$, then (5.42) is the feasible solution.

ii) If $T_3 \geq 0$ and $T_4 \geq 0$, then (5.43) is the feasible solution.

iii) If $T_2 \leq 0$ and $T_3 \leq 0$, then (5.44) is the feasible solution.

iv) If $T_1 \leq 0$ and $T_4 \leq 0$, then (5.45) is the feasible solution.

Proof of Theorem 5.6. Our first goal is to show that at most one of the four cases (5.42), (5.43), (5.44) and (5.45) leads to a feasible solution. To this end we note that (5.42) is feasible if and only if T_1 and T_2 are non-negative. Similar consideration from (5.43) to (5.45) leads to the Table 5.1.

Optimal Solution	T_1	T_2	T_3	T_4
(5.42) feasible	≥ 0	≥ 0	–	–
(5.43) feasible	–	–	≥ 0	≥ 0
(5.44) feasible	–	≤ 0	≤ 0	–
(5.45) feasible	≤ 0	–	–	≤ 0

Table 5.1: Behavior of expressions T_1, T_2, T_3 and T_4 when any optimal solution is feasible

Let us assume that (5.42) is feasible. By Lemma 5.5 we then see that (5.43) is not feasible. Since T_1 and T_2 are nonnegative, it is clear by Table 5.1 that (5.44) and (5.45) are not feasible. Hence, we have shown that if (5.42) is feasible, the remaining cases (5.43) to (5.45) are not feasible. Since analogous arguments can be repeated using Table 5.1 when one of the remaining cases (5.43) to (5.45) is considered feasible, we finally conclude that at most one of the four cases is feasible, that is, we have shown our intermediate result.

Let us now assume that none of the four cases yield a feasible solution. Then we obtain Table 5.2, where in each row, at least one of the inequalities needs to be true. Let us assume

Optimal Solution	T_1	T_2	T_3	T_4
(5.42) not feasible	< 0	< 0	–	–
(5.43) not feasible	–	–	< 0	< 0
(5.44) not feasible	–	> 0	> 0	–
(5.45) not feasible	> 0	–	–	> 0

Table 5.2: Behavior of expressions T_1, T_2, T_3 and T_4 when none of the optimal solutions is feasible

that $T_1 < 0$, then by Table 5.2, we conclude that we have following set of inequalities

$$kc_j < b_2 c_i \,, \tag{5.50}$$

$$kc_i > b_2 c_j \,, \tag{5.51}$$

$$b_1 c_i < kc_j \,, \tag{5.52}$$

$$b_1 c_j > kc_i \,. \tag{5.53}$$

Combining (5.50) and (5.52) as well as (5.51) and (5.53), we obtain

$$b_1 c_i < b_2 c_i \,, \tag{5.54}$$

$$b_2 c_j \quad < \quad b_1 c_j \,. \tag{5.55}$$

Now if $c_i < 0$, we find $c_j < 0$ by (5.50). Moreover (5.54) together with $c_i < 0$ implies $b_2 < b_1$, while (5.55) together with $c_j < 0$ implies $b_1 < b_2$, that is, we have found a contradiction. Analogously, we obtain a contradiction in the case $T_1 \geq 0$. As a consequence exactly one of the four cases produces a feasible solution. Finally, the implications are a direct consequence of the form of the solutions in (5.42) to (5.45) and the fact that only one case provides a feasible solution. ■

Theorem 5.6 also suggests to impose *if* conditions based on expressions (5.46), (5.47), (5.48)

Procedure 5 Compute feasible optimal 2D-solution

choose direction i and direction j

compute T_1, T_2, T_3 and T_4

if $T_1 \geq 0$ and $T_2 \geq 0$ **then**

$\alpha_i = 0$

$\beta_i = T_1/|B_1|$

$\alpha_j = 0$

$\beta_j = T_2/|B_1|$

else if $T_3 \geq 0$ and $T_4 \geq 0$ **then**

$\alpha_i = T_3/|B_2|$

$\beta_i = 0$

$\alpha_j = T_4/|B_2|$

$\beta_j = 0$

else if $T_2 < 0$ and $T_3 < 0$ **then**

$\alpha_i = 0$

$\beta_i = T3/|B_3|$

$\alpha_j = T_4/|B_3|$

$\beta_j = 0$

else

$\alpha_i = T1/|B_3|$

$\beta_i = 0$

$\alpha_j = 0$

$\beta_j = T_4/|B_3|$

end if

and (5.49) in the implementation of the algorithm for 2D-SVM solver, which is described in Procedure 5. This helps to reach directly to the feasible optimum solution.

5.3.2 Working Set Selection Strategies

From Procedure 5, we note that we need to choose the directions i^* and j^* in which the 2D-SVM solver performs an update. Several possibilities are available for this task. A straightforward approach is to consider all pairs of directions (i, j) and choose the one for which the 2D-gain of $\mathcal{D}$ is maximum. Note that the 2D-gain is simply an extension of the idea presented in Lemma 5.3. To be more precise, for $\alpha, \beta \in \mathbb{R}^n$ and $\delta, \eta \in \mathbb{R}^2$, the 2D-gain is

$$\mathcal{D}(\alpha+\delta_i e_i+\delta_j e_j, \beta+\eta_i e_i+\eta_j e_j)-\mathcal{D}(\alpha,\beta) = G(\delta_i,\eta_i)+G(\delta_j,\eta_j)-(\delta_i-\eta_i)(\delta_j-\eta_j)K_{i,j}\,, \quad (5.56)$$

where $G(\delta_k, \eta_k)$ for $k = i, j$ is the 1D-gain defined in Lemma 5.3.

It is worth noting that above described working set selection strategy is not a suitable choice because the search is $O(n^2)$. However it may be considered as an "optimal" two dimensional strategy and can be taken as a baseline to all other low-cost approximations to this approach. In the following, we describe two low-cost working set selection strategies following Steinwart et al (2011).

WSS 1: Two 1D-direction With Maximal Gain From Separate Subsets. In this approach, we preserve the low-cost search from 1D-solver. For this, we split the index set $\{1, \ldots, n\}$ into two parts $\{1, \ldots, \frac{n}{2}\}$ and $\{\frac{n}{2}+1, \ldots, n\}$ and search for 1D directions with maximum gain over these two parts separately. In other words, we can choose the directions i^* and j^* by

$$\begin{aligned} i^* &\in \arg\max_{i \le n/2} \mathcal{D}(\alpha+\delta e_i, \beta+\eta e_i) - \mathcal{D}(\alpha,\beta)\,, \\ j^* &\in \arg\max_{i > n/2} \mathcal{D}(\alpha+\delta e_i, \beta+\eta e_i) - \mathcal{D}(\alpha,\beta)\,, \end{aligned} \quad (5.57)$$

where δ and η are defined in the 1D-solution. These chosen directions are used for the first iteration. For the subsequent iterations, we first search for two new 1D directions, i^*_{new} and j^*_{new}, using again by (5.57). Then we compute the 2D-gain of $\mathcal{D}$ *for all pairs* of old and new choosen directions of the previous and the current iterations, respectively, and choose the pair for which the 2D-gain is maximum.

WSS 2: 1D-direction With Maximal Gain And A Direction Of A Nearby Sample. This strategy is simply an extension of *WSS 1*. After determining (i^*, j^*) by *WSS 1*, we fix i^* and then search for another direction j^* from k-nearest neighbors of x_{i^*} with respect to the metric

$$d(x, x') := \|x - x'\|^2\,,$$

for which $2D$-gain is maximum as compared to the j^* chosen by *WSS 1*.

5.3.3 Stopping Criteria

To formulate the stopping criterion for the 2D-problem, we follow the idea that is presented in Section 3.1. Let us first consider the component $T(\alpha, \beta)$ from (5.26) and by using (5.56), we find the following update of $T(\alpha, \beta)$ in the directions of i and j

$$T(\alpha + \delta e_i + \delta e_j, \beta + \eta e_i + \eta e_j) = T(\alpha, \beta) - U(\alpha_i, \beta_i, \delta_i, \eta_i) - U(\alpha_j, \beta_j, \delta_j, \eta_j) + 2(\delta_i - \eta_i)(\delta_j - \eta_j)K_{i,j}\,,$$

where $U(\alpha_k, \beta_k, \delta_k, \eta_k)$ for $k = i, j$ is defined in (5.27). To compute $E(\alpha, \beta)$, we first obtain the updated gradients in the directions of i and j, and then subsequently compute $\xi_{l,+}, \xi_{l,-}$. Moreover, $\widehat{E}(\alpha, \beta)$ can also be computed for the 2D-problem similar to the 1D-problem by using (5.30). With all above computations, we summarize 2D-SVM solver in Algorithm 2.

Algorithm 2 2D-SVM Solver

initialize $\alpha, \beta, \nabla\mathcal{D}_\alpha(\alpha, \beta), \nabla\mathcal{D}_\beta(\alpha, \beta)$ and $T(\alpha, \beta)$

while $S(\alpha, \gamma) > \frac{\varepsilon}{2\alpha}$ **do**

select directions i^* and j^*

use procedure 5 to obtain the optimum solution for direction i^* and j^*

update α and β in the direction i^* and j^*

update $\nabla\mathcal{D}_\alpha(\alpha, \beta), \nabla\mathcal{D}_\beta(\alpha, \beta)$ in the directions (i^*, j^*) and calculate $S(\alpha, \beta)$

end while

5.4 Convergence Analysis

In this section, we show that the 1D-SVM and the 2D-SVM converge to the optimal solution. We start by following lemma.

Lemma 5.7. *For all $\rho \in (-\infty, D^*)$, the set $\{(\alpha, \beta) : \mathcal{D}(\alpha, \beta) \geq \rho\}$ is compact.*

Proof of Lemma 5.7. Since the kernel matrix K in (5.7) is positive definite, we find

$$\begin{aligned}\mathcal{D}(\alpha, \beta) &\leq \langle \alpha - \beta, y\rangle - \frac{1}{4C\tau}\langle \alpha, \alpha\rangle - \frac{1}{4C(1-\tau)}\langle \beta, \beta\rangle \\ &\leq \|\alpha - \beta\|\|y\| - \frac{1}{4C\tau}\|\alpha\|^2 - \frac{1}{4C(1-\tau)}\|\beta\|^2 \\ &\leq \|\alpha - \beta\|\|y\| - \frac{1}{4C}\|\alpha\|^2 - \frac{1}{4C}\|\beta\|^2\end{aligned}$$

$$= \frac{1}{4C}\Big(\tilde{C}\|\alpha-\beta\| - \|\alpha\|^2 - \|\beta\|^2\Big)\,,$$

where $\tilde{C} := 4C\|y\|$. For $\alpha \geq 0$ and $\beta \geq 0$ with $\mathcal{D}(\alpha,\beta) \geq \rho$ we thus have

$$\tilde{C}\|\alpha-\beta\| - \|\alpha\|^2 - \|\beta\|^2 \geq 4C\rho\,,$$

which implies that

$$\|\alpha\|^2 + \|\beta\|^2 \leq \tilde{C}\|\alpha-\beta\| - 4C\rho \leq \tilde{C}\Big(\|\alpha\| + \|\beta\|\Big) - 4C\rho\,. \tag{5.58}$$

From (5.58), we easily conclude that the set $\{(\alpha,\beta) : \mathcal{D}(\alpha,\beta) \geq \rho\}$ is bounded. Furthermore, the set $\{(\alpha,\beta) : \mathcal{D}(\alpha,\beta) \geq \rho\}$ is also closed, since $\mathcal{D}(\alpha,\beta)$ is continuous. Thus, the set is compact. ■

Lemma 5.7 implies that for $k \in \mathbb{N}$ the sequences $\{\alpha^k\}$ and $\{\beta^k\}$ produced by algorithm lie in $\{(\alpha,\beta) : \mathcal{D}(\alpha,\beta) \geq \rho\}$ and are bounded. In other words, $\alpha \in [0,A]^n$ and $\beta \in [0,B]^n$, where $A \geq 0$ and $B \geq 0$. Let us define sigma functional for vectors α and β for an index set $I \subset \{1,\ldots,n\}$, which represents first-order approximation of the maximal distance (with regard to the value of the dual objective function) between a given solution and any other feasible solution.

$$\sigma(\alpha,\beta|I) := \sup_{\substack{\tilde{\alpha}\in[0,A],\tilde{\beta}\in[0,B]\\ \tilde{\alpha}_i=\alpha_i,\tilde{\beta}_i=\beta_i\forall i\neq I}} \Big(\langle\nabla\mathcal{D}_\alpha(\alpha,\beta),\tilde{\alpha}-\alpha\rangle + \langle\nabla\mathcal{D}_\beta(\alpha,\beta),\tilde{\beta}-\beta\rangle\Big)\,. \tag{5.59}$$

Since our algorithms are based on gain optimization, we further define the γ-functional

$$\gamma(\alpha,\beta|I) := \sup_{\substack{\tilde{\alpha}\in[0,\infty),\tilde{\beta}\in[0,\infty)\\ \tilde{\alpha}_i=\alpha_i,\tilde{\beta}_i=\beta_i\forall i\neq I}} \mathcal{D}(\tilde{\alpha},\tilde{\beta}) - \mathcal{D}(\alpha,\beta)\,, \tag{5.60}$$

which is the gain in dual objective function resulting from an optimization over the directions contained in I. To simplify notations, we write $\sigma(\alpha,\beta|i) := \sigma(\alpha,\beta|\{i\})$ and $\gamma(\alpha,\beta|i) := \gamma(\alpha,\beta|\{i\})$, that is

$$\sigma(\alpha,\beta|i) = \sup_{\tilde{\alpha}_i\in[0,A],\tilde{\beta}_i\in[0,B]} \Big((\tilde{\alpha}_i-\alpha_i)\nabla\mathcal{D}_{\alpha_i}(\alpha,\beta) + (\tilde{\beta}_i-\beta_i)\nabla\mathcal{D}_{\beta_i}(\alpha,\beta)\Big)\,,$$

and the gain of the 1D-update in the direction i is

$$\gamma(\alpha,\beta|i) = \sup_{\tilde{\alpha}_i\in[0,\infty),\tilde{\beta}_i\in[0,\infty)} \mathcal{D}(\alpha+\delta_i,\beta_i+\eta_i)) - \mathcal{D}(\alpha,\beta)\,,$$

where $\delta_i := \alpha_i^{\text{new}} - \alpha_i$ and $\eta_i := \beta_i^{\text{new}} - \beta_i$. Moreover, for $I = \{1,,\ldots,n\}$, we write $\sigma(\alpha,\beta) := \sigma(\alpha,\beta|I)$ and $\gamma(\alpha,\beta) := \gamma(\alpha,\beta|I)$ respectively. Note that both σ and γ are monotonic in I, that is, for $I \subset J$, we have $\sigma(\alpha,\beta|I) \leq \sigma(\alpha,\beta|J)$ and $\gamma(\alpha,\beta|I) \leq \gamma(\alpha,\beta|J)$.

In the following lemma we establishes a relationship between the sigma functional and the gamma functional for $I = \{1,\ldots,n\}$.

Lemma 5.8. *For all $\alpha \in [0,A]^n$ and $\beta \in [0,B]^n$, we have*

$$\sum_{i=1}^{n} \sigma(\alpha,\beta|i) = \sigma(\alpha,\beta) \geq \gamma(\alpha,\beta)\,.$$

In particular, there exists an index $i^\star \in \{1,\dots,n\}$ such that

$$\sigma(\alpha,\beta|i^\star) \geq n^{-1}\sigma(\alpha,\beta)\,. \tag{5.61}$$

Proof of Lemma 5.8. For $i \in \{1,\dots,n\}$, we define

$$\bar\alpha_i := \begin{cases} A & \text{if } \nabla\mathcal{D}_{\alpha_i}(\alpha,\beta) \geq 0 \\ 0 & \text{if } \nabla\mathcal{D}_{\alpha_i}(\alpha,\beta) < 0 \end{cases} \quad \text{and} \quad \bar\beta_i := \begin{cases} B & \text{if } \nabla\mathcal{D}_{\beta_i}(\alpha,\beta) \geq 0 \\ 0 & \text{if } \nabla\mathcal{D}_{\beta_i}(\alpha,\beta) < 0\,. \end{cases} \tag{5.62}$$

Then the vectors $\bar\alpha := (\bar\alpha_1,\dots,\bar\alpha_n) \in [0,A]^n$ and $\bar\beta := (\bar\beta_1,\dots,\bar\beta_n) \in [0,B]^n$ realize the supremum defining $\sigma(\alpha,\beta)$, and hence we obtain

$$\begin{aligned}\sum_{i=1}^{n} \sigma(\alpha,\beta|i) &= \sum_{i=1}^{n} \Big(\langle \nabla\mathcal{D}_\alpha(\alpha,\beta), (\bar\alpha_i - \alpha_i)e_i\rangle + \langle \nabla\mathcal{D}_\beta(\alpha,\beta), (\bar\beta_i - \beta_i)e_i\rangle \Big) \\ &= \langle \nabla\mathcal{D}_\alpha(\alpha,\beta), (\bar\alpha - \alpha)\rangle + \langle \nabla\mathcal{D}_\beta(\alpha,\beta), (\bar\beta - \beta)\rangle \\ &= \sigma(\alpha,\beta)\,.\end{aligned}$$

To show the inequality $\sigma(\alpha,\beta) \geq \gamma(\alpha,\beta)$, we fix an $\tilde\alpha \in [0,A]^n$ and a $\tilde\beta \in [0,B]^n$ for $I = \{1,\dots,n\}$. Additionally, we write $b_3 := \frac{1}{4C\tau}$ and $b_4 := \frac{1}{4C(1-\tau)}$. Then we find

$$\begin{aligned}&\mathcal{D}(\tilde\alpha,\tilde\beta) - \mathcal{D}(\alpha,\beta) \\ &= \langle \tilde\alpha - \tilde\beta, y\rangle - \frac{1}{2}\langle \tilde\alpha - \tilde\beta, K(\tilde\alpha - \tilde\beta)\rangle - b_3\langle\tilde\alpha,\tilde\alpha\rangle - b_4\langle\tilde\beta,\tilde\beta\rangle \\ &\quad - \langle \alpha - \beta, y\rangle + \frac{1}{2}\langle \alpha - \beta, K(\alpha - \beta)\rangle + b_3\langle\alpha,\alpha\rangle + b_4\langle\beta,\beta\rangle \\ &= \langle \tilde\alpha - \alpha, y\rangle - \langle \tilde\alpha - \alpha, K(\alpha - \beta)\rangle - 2b_3\langle\tilde\alpha - \alpha,\alpha\rangle - \langle \tilde\beta - \beta, y\rangle + \langle \tilde\beta - \beta, K(\alpha - \beta)\rangle \\ &\quad - 2b_4\langle\tilde\beta - \beta,\beta\rangle + \langle \tilde\alpha - \alpha, K(\alpha - \beta)\rangle - \langle \tilde\beta - \beta, K(\alpha - \beta)\rangle - \frac{1}{2}\langle \tilde\alpha - \tilde\beta, K(\tilde\alpha - \tilde\beta)\rangle \\ &\quad + \frac{1}{2}\langle \alpha - \beta, K(\alpha - \beta)\rangle + 2b_3\langle\tilde\alpha - \alpha,\alpha\rangle - b_3\langle\tilde\alpha,\tilde\alpha\rangle + b_3\langle\alpha,\alpha\rangle + 2b_4\langle\tilde\beta - \beta,\beta\rangle \\ &\quad - b_4\langle\tilde\beta,\tilde\beta\rangle + b_4\langle\beta,\beta\rangle\,.\end{aligned}$$

A simple calculations shows that

$$\begin{aligned}&\langle \tilde\alpha - \alpha, K(\alpha - \beta)\rangle - \langle \tilde\beta - \beta, K(\alpha - \beta)\rangle - \frac{1}{2}\langle \tilde\alpha - \tilde\beta, K(\tilde\alpha - \tilde\beta)\rangle + \frac{1}{2}\langle \alpha - \beta, K(\alpha - \beta)\rangle \\ &= \langle \tilde\alpha, K\alpha\rangle - \langle \tilde\alpha, K\beta\rangle - \langle \alpha, K\alpha\rangle + \langle \alpha, K\beta\rangle - \langle \tilde\beta, K\alpha\rangle + \langle \tilde\beta, K\beta\rangle + \langle \beta, K\alpha\rangle - \langle \beta, K\beta\rangle \\ &\quad - \frac{1}{2}\langle \tilde\alpha, K\tilde\alpha\rangle + \langle \tilde\alpha, K\tilde\beta\rangle - \frac{1}{2}\langle \tilde\beta, K\tilde\beta\rangle + \frac{1}{2}\langle \alpha, K\alpha\rangle - \langle \alpha, K\beta\rangle + \frac{1}{2}\langle \beta, K\beta\rangle\end{aligned}$$

$$
\begin{aligned}
&= \Big(\langle\tilde{\alpha}, K\alpha\rangle - \frac{1}{2}\langle\alpha, K\alpha\rangle - \frac{1}{2}\langle\tilde{\alpha}, K\tilde{\alpha}\rangle\Big) + \Big(\langle\tilde{\beta}, K\beta\rangle - \frac{1}{2}\langle\beta, K\beta\rangle - \frac{1}{2}\langle\tilde{\beta}, K\tilde{\beta}\rangle\Big) \\
&\quad + \Big(\langle\alpha, K\beta\rangle - \langle\tilde{\alpha}, K\beta\rangle - \langle\tilde{\beta}, K\alpha\rangle + \langle\tilde{\alpha}, K\tilde{\beta}\rangle\Big) \\
&= -\frac{1}{2}\langle\tilde{\alpha} - \alpha, K(\tilde{\alpha} - \alpha)\rangle - \frac{1}{2}\langle\tilde{\beta} - \beta, K(\tilde{\beta} - \beta)\rangle + \langle\tilde{\alpha} - \alpha, K(\tilde{\beta} - \beta)\rangle \\
&= -\frac{1}{2}\langle\tilde{\alpha} - \alpha - \tilde{\beta} + \beta, K(\tilde{\alpha} - \alpha - \tilde{\beta} + \beta)\rangle\,.
\end{aligned}
$$

Thus, by using the gradients of α and β for $I = \{1, \ldots, n\}$, see (5.9), we obtain

$$
\begin{aligned}
&\mathcal{D}(\tilde{\alpha}, \tilde{\beta}) - \mathcal{D}(\alpha, \beta) \\
&= \langle\nabla\mathcal{D}_\alpha(\alpha, \beta), \tilde{\alpha} - \alpha\rangle + \langle\nabla\mathcal{D}_\beta(\alpha, \beta), \tilde{\beta} - \beta\rangle - \frac{1}{2}\langle\tilde{\alpha} - \alpha - \tilde{\beta} + \beta, K(\tilde{\alpha} - \alpha - \tilde{\beta} + \beta)\rangle \\
&\quad - b_3\langle\tilde{\alpha} - \alpha, \tilde{\alpha} - \alpha\rangle - b_4\langle\tilde{\beta} - \beta, \tilde{\beta} - \beta\rangle \\
&\leq \langle\nabla\mathcal{D}_\alpha(\alpha, \beta), \tilde{\alpha} - \alpha\rangle + \langle\nabla\mathcal{D}_\beta(\alpha, \beta), \tilde{\beta} - \beta\rangle\,,
\end{aligned}
$$

and maximizing on both sides of this inequality, we find $\boldsymbol{\gamma}(\alpha, \beta) \leq \sigma(\alpha, \beta)$. Finally, the last assertion is a trivial consequence of the first assertion. ■

Now we present the lemma that relates the sigma functional $\sigma(\alpha, \beta|i)$ to the gain, i.e. gamma functional $\boldsymbol{\gamma}(\alpha, \beta|i)$ for direction $i \in \{1, \ldots, n\}$.

Lemma 5.9. *For all $\alpha \in [0, A]^n$, $\beta \in [0, B]^n$ and $i \in \{1, \ldots, n\}$, we have*

$$
\boldsymbol{\gamma}(\alpha, \beta|i) \geq \frac{\sigma(\alpha, \beta|i)}{2} \min\left\{1, \frac{\sigma(\alpha, \beta|i)}{M}\right\}\,,
$$

where b_1 and b_2 are defined in (5.12), and $M := b_1A^2 + b_2B^2$.

Proof of Lemma 5.9. Let $\bar{\alpha}_i$ and $\bar{\beta}_i$ be defined by (5.62), and $d_1 := \bar{\alpha}_i - \alpha_i$ and $d_2 := \bar{\beta}_i - \beta_i$. Then for $\lambda \in [0, 1]$, we have $\alpha + \lambda d_1 \in [0, A]$ and $\beta + \lambda d_2 \in [0, B]$ respectively. Furthermore, we define $\delta := \lambda d_1$ and $\eta := \lambda d_2$. Then by using the 1D gain of the dual objective function, see (5.22), together with (5.59) for $i \in \{1, , \ldots, n\}$, we obtain

$$
\begin{aligned}
&\mathcal{D}(\alpha + \delta e_i, \beta + \eta e_i) - \mathcal{D}(\alpha, \beta) \\
&= \lambda d_1\Big(\nabla\mathcal{D}_{\alpha_i}(\alpha, \beta) - \frac{\lambda b_1 d_1}{2}\Big) + \lambda d_2\Big(\nabla\mathcal{D}_{\beta_i}(\alpha, \beta) - \frac{\lambda b_2 d_2}{2}\Big) + \lambda^2 d_1 d_2 \\
&= \lambda\Big(d_1\nabla\mathcal{D}_{\alpha_i}(\alpha, \beta) + d_2\nabla\mathcal{D}_{\beta_i}(\alpha, \beta)\Big) - \frac{\lambda^2}{2}\Big(b_1 d_1^2 - 2d_1 d_2 + b_2 d_2^2\Big) \\
&\geq \lambda\sigma(\alpha, \beta|i) - \frac{\lambda^2}{2}\Big(b_1 A^2 + b_2 B^2\Big)\,. \qquad (5.63)
\end{aligned}
$$

In order to maximize the right hand side of (5.63), we consider

$$
h(\lambda) := \lambda\sigma(\alpha, \beta|i) - \frac{\lambda^2}{2}M\,,
$$

where $M := b_1 A^2 + b_2 B^2$. Simple calculations show that $h(\lambda)$ attains maximum value at

$$\lambda^* := \begin{cases} 1 & \text{if } \sigma(\alpha, \beta | i) > M \\ \frac{\sigma(\alpha,\beta|i)}{M} & \text{if } \sigma(\alpha, \beta | i) \leq M \,. \end{cases}$$

In the case $\sigma(\alpha, \beta|i) > M$, we then find

$$h(\lambda^*) = \sigma(\alpha, \beta | i) - \frac{M}{2} \geq \frac{\sigma(\alpha, \beta | i)}{2} \,,$$

while in the other case $\sigma(\alpha, \beta|i) \leq M$, we obtain

$$\gamma(\alpha, \beta | I) \geq \frac{\sigma^2(\alpha, \beta | i)}{2M} \,.$$

By combining the above two estimates, we achieve

$$h(\lambda^*) \geq \frac{\sigma(\alpha, \beta | i)}{2} \min \left\{ 1, \frac{\sigma(\alpha, \beta | i)}{M} \right\} .$$

Finally, by the definition of the gamma functional (5.60), we have

$$\gamma(\alpha, \beta | i) \geq \mathcal{D}(\alpha + \lambda^* d_1, \beta + \lambda^* d_2) - \mathcal{D}(\alpha, \beta) \geq \frac{\sigma(\alpha, \beta | i)}{2} \min \left\{ 1, \frac{\sigma(\alpha, \beta | i)}{M} \right\} .$$

■

With these preparations we now establish the result that leads to the convergence of both the 1D and the 2D solvers for SVMs.

Theorem 5.10. *Let $(\alpha^{(0)}, \beta^{(0)}), (\alpha^{(1)}, \beta^{(1)}), \cdots \in [0, A]^n \times [0, B]^n$ be a sequence of feasible vectors that satisfies*

$$\mathcal{D}(\alpha^{(\ell+1)}, \beta^{(\ell+1)}) - \mathcal{D}(\alpha^{(\ell)}, \beta^{(\ell)}) \geq \gamma(\alpha^{(\ell)}, \beta^{(\ell)} | i_\ell^\star) \,, \qquad \ell \geq 0 \,, \tag{5.64}$$

where for each ℓ, the index $i_\ell^\star \in \{1, \ldots, n\}$ satisfies (5.61). Then for $\ell \to \infty$, we have $\mathcal{D}(\alpha^{(\ell)}, \beta^{(\ell)}) \to D^*$, where $\mathcal{D}^*$ is the optimal value of $\mathcal{D}(\alpha, \beta)$.

Proof of Theorem 5.10. Since $\gamma(\alpha^{(\ell)}, \beta^{(\ell)} | i_\ell^\star) \geq 0$, the sequence $\mathcal{D}(\alpha^{(\ell)}, \beta^{(\ell)})$ is monotonically increasing and since it is bounded by $\mathcal{D}^*$, we see $\mathcal{D}(\alpha^{(\ell)}, \beta^{(\ell)})$ is a Cauchy sequence and thus $\mathcal{D}(\alpha^{(\ell+1)}, \beta^{(\ell+1)}) - \mathcal{D}(\alpha^{(\ell)}, \beta^{(\ell)}) \to 0$ when $\ell \to \infty$. This implies that $\gamma(\alpha^{(\ell)}, \beta^{(\ell)} | i_\ell^\star) \to 0$ by (5.64). By Lemma 5.9, this implies that $\sigma(\alpha^{(\ell)}, \beta^{(\ell)} | i_\ell^\star) \to 0$, which, by Lemma 5.8, leads to $\sigma(\alpha^\ell, \beta^\ell) \to 0$. This further implies that $\mathcal{D}^* - \mathcal{D}(\alpha^\ell, \beta^\ell) = \gamma(\alpha^\ell, \beta^\ell) \to 0$ by Lemma 5.8. ■

Note that the working set selection strategies WSS 1 and WSS 2 satisfy the assumption (5.64), since they both achieve a gain that is at least as large as the best 1D-gain. Furthermore, Theorem 5.10 leads to the following corollary showing that the duality gap vanishes for $\ell \to \infty$.

Corollary 5.11. *Under the assumptions of Theorem 5.10, for $\ell \to \infty$, we have*

$$\mathbf{w}^{(\ell)} := \sum_{i=1}^{n} (\alpha_i^{(\ell)}, \beta_i^{(\ell)}) \phi(\mathbf{x}_i) \to \mathbf{w}^* \,, \qquad \ell \geq 0 \,,$$

where $\mathbf{w}^*$ is defined by (5.8) . In particular, $S(\alpha^{(\ell)}, \beta^{(\ell)}) \to 0$, where $S(\alpha, \beta)$ is a duality gap defined by (5.25), and hence Algorithm 1 and Algorithm 2 terminate after finitely many iterations.

Proof of Corollary 5.11. Let $(\alpha^{(\ell)}, \beta^{(\ell)})$ be a sequence satisfying (5.64). Then by Theorem 5.10, we see that

$$\mathcal{D}(\alpha^{(\ell)}, \beta^{(\ell)}) \to \mathcal{D}^* \,. \tag{5.65}$$

Since $(\alpha^{(\ell)}, \beta^{(\ell)})$ is contained in the set $\mathcal{A} := \{(\alpha, \beta) : \mathcal{D}(\alpha, \beta) \geq \mathcal{D}(\alpha^{(0)}, \beta^{(0)})\}$, which is bounded by Lemma 5.7, there thus exists an (α^*, β^*) and a subsequence $(\alpha^{(\ell_k)}, \beta^{(\ell_k)})$ with $(\alpha^{(\ell_k)}, \beta^{(\ell_k)}) \to (\alpha^*, \beta^*)$. From (5.65) we conclude that $\mathcal{D}(\alpha^*, \beta^*) = \mathcal{D}^*$. Let us now consider $\mathbf{w}^{(\ell)}$ and $\mathbf{w}^*$. Since (α^*, β^*) is a solution of the dual problem, $(\mathbf{w}^*, \xi_+^*, \xi_-^*)$ is the solution of corresponding primal problem and by (Steinwart and Christmann, 2008, Lemma 5.1), this primal solution is unique. Moreover $(\alpha^{(\ell_k)}, \beta^{(\ell_k)}) \to (\alpha^*, \beta^*)$ shows that

$$\mathbf{w}^{(\ell_k)} \to \mathbf{w}^* \,.$$

It remains to show that this convergence holds for the sequence $\mathbf{w}^{(\ell)}$, too. Let us assume the converse, that is, there exists a subsequence $\mathbf{w}^{(\ell'_k)}$ and an $\varepsilon > 0$ with

$$\|\mathbf{w}^{(\ell'_k)} - \mathbf{w}^*\|_H \geq \varepsilon \,. \tag{5.66}$$

By the compactness of set $\mathcal{A}$, this subsequence yields a sub-subsequence $(\alpha^{(\ell''_k)}, \beta^{(\ell''_k)})$ of $(\alpha^{(\ell'_k)}, \beta^{(\ell'_k)})$ and $(\alpha^{(\ell^{**}_k)}, \beta^{(\ell^{**}_k)})$ with

$$(\alpha^{(\ell''_k)}, \beta^{(\ell''_k)}) \to (\alpha^{(\ell^{**}_k)}, \beta^{(\ell^{**}_k)}) \,.$$

Repeating aforementioned arguments for

$$\mathbf{w}^{**} := \sum_{i=1}^{n} (\alpha_i^{**} - \beta_i^{**}) K(x_i, \cdot) \,,$$

we find $\mathbf{w}^{(\ell''_k)} \to \mathbf{w}^{**}$. However, since $\mathbf{w}^{**}$ together with its corresponding optional slack variables ξ_+^{**}, ξ_-^{**} form a solution of primal problem, we have $\mathbf{w}^{**} = \mathbf{w}^*$ by the uniqueness of the primal solution. We have found a contradiction to (5.66), and consequently, we have

$$\mathbf{w}^{(\ell)} \to \mathbf{w}^* \,.$$

In other words

$$\xi_{i,+}^{(\ell)} := \max\left\{0, y_i - \mathbf{w}^{(\ell)}(x_i)\right\} \to \xi_{i,+}^*$$
$$\xi_{i,-}^{(\ell)} := \max\left\{0, -y_i + \mathbf{w}^{(\ell)}(x_i)\right\} \to \xi_{i,-}^*\,,$$

and thus we find

$$\mathcal{P}_C(\mathbf{w}^{(\ell)}, \xi_{i,+}^{(\ell)}, \xi_{i,-}^{(\ell)}) \to \mathcal{P}_C(\mathbf{w}^*, \xi_+^*, \xi_-^*)\,.$$

Since $\mathcal{P}_C(\mathbf{w}^*, \xi_{i,+}^*, \xi_{i,-}^*) = \mathcal{D}^*$, we finally have $S(\alpha^{(\ell)}, \beta^{(\ell)}) \to 0$ and thus the stopping criteria of Algorithm 1 and Algorithm 2 are satisfied after finitely many iterations. ■

5.5 Experiments

To evaluate the performance of the proposed solver for expectile regression, we perform several experiments to address the following questions:

1. Which subset selection strategy leads to the smallest number of iterations or shortest run time?

2. What is the number of nearest neighbors that leads to the smallest number of iterations and shortest run time?

3. Is there advantage of warm start initialization when the parameter search is performed over a grid?

4. Does the clipping provide a significant reduction in the training time and iterations?

5. How well does the 2D-SVM-solver work as compared to ER-Boost that is proposed by Yang and Zou (2015)?

6. Does expectile regression give a computational advantage over qunatile regression?

To answer these questions, we implemented the 2D-SVM-solver in C^{++}, where the source code (ex-svm) is now a part of the larger package *liquidSVM: A Fast and Versatile SVM Package*, and can be downloaded from `http://www.isa.uni-stuttgart.de/software/`, see also Steinwart and Thomann (2017). The algorithm was compiled by LINUX's gcc version 4.7.2 with various software and hardware optimization flags enabled. All experiments were conducted on a computer with INTEL CORE i7-4770 (3.40 GHz) and 16GB RAM under 64bit version

of Debian 3.16.0-4-amd64. During all experiments that incorporated the measurement of run time, one core was used solely for the experiments, and the number of other processes running on the system were minimized.

In order to perform the experiments, we considered nine data sets that were downloaded from different sources. These data sets comprises various number of features and vary in sample sizes from 630 to 20639. The data sets CONCRETE-COMP, UPDRS-MOTOR, CYCLE-PP, AIRFOIL-NOISE and HOUR were downloaded from UCI repository. The two data sets NC-CRIME and HEAD-CIRCUM are available and documented in R packages Ecdat and AGD respectively. The remaining two data sets CAL-HOUSING and MUNICH-RENT were downloaded from StatLib and the data archive of the Institute of Statistics, Ludwig-Maximilians-University of Munich respectively. We scaled the data sets componentwise such that all the samples including labels lie in $[-1,1]^{d+1}$, where d represents dimensions of the input data. In addition to that, we generated a random split for all data sets that contained approximately 70% training and 30% test samples. The characteristics of the considered data sets are described in Table 5.3.

In all our experiments with SVM solver, we used the Gaussian kernel (5.3). To determine the hyper-parameters, we considered a geometrically spaced 10 by 10 grid for λ and γ over the interval $[c_1 n^{-1}, 1]$ and $[c_2 n^{-1/d}, c_3]$ respectively, where n is the number of training samples, d is the input dimension, and $c_1 := 0.001, c_2 := 0.1$ and $c_3 := 0.2$. Here, the values of the constants were chosen with the help of our experience, see also Chapter 4. To choose the best values

data	sample sizes	training size	test size	dimension
NC-CRIME	630	441	189	19
CONCRETE-COMP	1030	721	309	8
AIRFOIL-NOISE	1503	1052	451	5
MUNICH-RENT	2053	1437	616	12
UPDRS-MOTOR	5875	4112	1763	19
HEAD-CIRCUM	7020	4914	2106	4
CYCLE-PP	9568	6697	2871	5
HOUR	17379	12165	5214	12
CAL-HOUSING	20639	14447	6192	8

Table 5.3: Characteristics of data sets together with the training sizes and the test sizes that refer to the splits used in the run time experiments.

of these hyper-parameters, we used k-fold cross validation with randomly generated folds. In our case, we considered $k = 5$. During the k-fold cross validation, the hyper-parameter γ was internally converted to $\tilde{\gamma} := \frac{(k-1)n\gamma}{k}$ and λ to $C := \frac{k}{2(k-1)n\lambda}$, where $(k-1)n/k$ is approximately the *actual* training set size for k-fold cross validation. Finally, we have performed the experiments for each $\tau = 0.25, 0.50, 0.75$ to investigate the performance of the solver based on all the aforementioned questions.

Let us now explore the answers of the above stated questions one by one. To address the first question, we performed experiments with warm start initialization method and clipped duality gap. In addition, we considered $N = 15$ nearest neighbors for WSS 2. The results are presented in Figure A.1 and A.2 (see, Appendix A for all figures A.x), which depict that WSS 2 needs substantially less iterations as well as training time than WSS 1 on all data sets. For larger data sets such as UPDRS-MOTOR, HEAD-CIRCUM, CYCLE-PP, HOURS and CAL-HOUSING, the run time and iterations with WSS 2 is at least 50% less than WSS 1. Moreover, a closer analysis, see Figure A.3 and A.4, shows that the savings are obtained at the hyper-parameters pairs for which training is particularly expensive, that is, for small λ and medium to small γ.

Note that we have fixed $N = 15$ for WSS 2 to address the previous question. To investigate how the computational requirements change with the number of nearest neighbors, we performed the experiments for N-nearest neighbors with $N = 5, 10, 15, 20, 25, 30, 35, 40$. Again we used the warm start initialization and the clipped duality gap. Here, it was observed that the number of iterations tends to decrease with increasing N. However, for $N \geq 25$, only a slight improvement in the number of iterations was found whereas the required run time tended to increase compared to smaller N. We therefore plotted the results for $N = 5, 10, 15, 20$ only. Figure A.5 shows that the solver attains the minimum training time for $N = 15$ on almost all data sets. Moreover, Figure A.6 shows that the number of iterations decreases with increasing N. However, this decrease becomes negligible when $N \geq 15$. All this together leads us to conclude that $N = 15$ is the best choice for our EX-SVM solver. Finally, Figure A.7 and A.8 illustrate the computational requirements for different hyper-parameters pairs. Again the largest savings for $N = 15$ were obtained for small λ.

To answer the third question regarding the initialization methods, we trained with $N = 15$ and the clipped duality gap. The results, which are presented in Figure A.9 and A.10 show that using of the warm start initialization saves between 20% and 40% of both training time and iterations. The detailed behavior for different hyper-parameter pairs is illustrated in Figure A.11 and A.12. Again the savings are more pronounced for smaller λ.

To answer the forth question, we considered the stopping criterion with clipped and with unclipped duality gap. Here, we used the warm start initialization and WSS 2 with $N = 15$ nearest neighbors. The corresponding results are shown in Figures A.13 and A.14. In the case of the hinge loss function, Steinwart et al (2011) showed that using the clipped duality gap yields significant reduction, both in run time and the number of iterations. In our case, however, we get only a small reduction in iterations, that is, 1% on almost all data sets. On the other hand, this stopping criterion causes 2% to 17% increase in run times on the data sets. This indicates that the unclipped duality gap is the better choice in our case. The per grid plot of hyper-parameters for data set CAL-HOUSING, as presented in Figure A.15 and A.16, shows that clipping reduces the run time only for few pairs of hyper-parameter when λ is small and γ is large. For rest of the pairs, unclipped duality gap leads to smaller run time.

Finally, to answer the last two questions, we compare the preformance of our SVM solver with *a)* ER-BOOST on the basis of test error and *b)* with both ER-BOOST and a similarly implemented SVM solver for quantile regression on the basis of training time, see Steinwart and Thomann (2017) for the source code.

For this, we considered the 2D-SVM solver with unclipped duality gap for expectile (EX-SVM) and quantile (Q-SVM), the 2D-SVM solver with clipped duality gap for expectile (EX-SVM*) and quantile (Q-SVM*), and ER-BOOST, see (Yang and Zou, 2015). Since the experiments using large data sets entail long run times, we splitted the data sets into three categories, namely, small ($n < 5000$), medium ($5000 \leq n < 10000$) and large ($n \geq 10000$). We then conducted experiments for EX-SVM, EX-SVM*, Q-SVM, Q-SVM* and ER-BOOST by repeating 5-fold cross validation 25, 10 and 5 times for the small, medium and large data sets respectively. For 2D-SVM solvers, we used the 10 by 10 default grid of hyper-parameters as described above. For ER-BOOST, we used the default value of boosting steps ($M = 100$) and performed 5-fold cross validation to choose the best value of the interaction level (L) between variables, as described by the ER-BOOST manual. The resulting, average test errors (standard deviation) and training times are shown in Table 5.4 and Table 5.5 respectively. It turns out that both SVM solvers exhibit a better test performance than ER-BOOST on all data sets, but all reported errors are relatively small. Examining the achieved training times for each data set, we observe that SVM solvers, both for expectile and quantile, are more sensitive to the training set size and less sensitive to the dimensions of data set, whereas, ER-BOOST behaves the other way around. In addition to that, the test performance of EX-SVM* is slightly better than EX-SVM at the cost of almost 10% longer training times. In addition, we see that the expectile solver is, depending

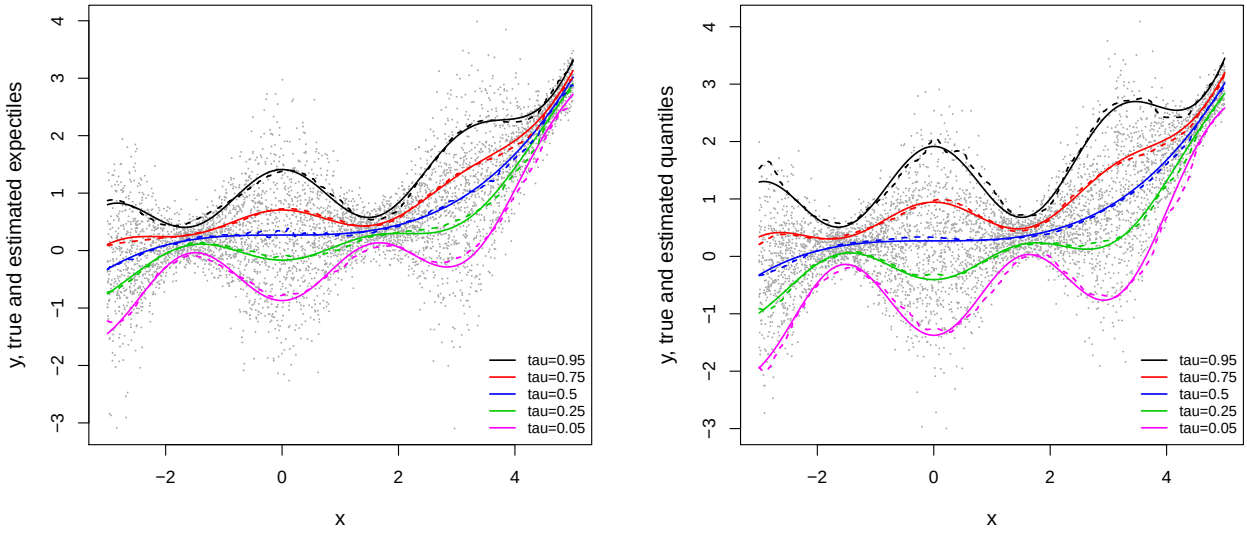

Figure 5.1: True (solid) and estimated (dashed) expectiles (left) and corresponding quantiles (right) for $\tau = 0.05, 0.25, 0.5, 0.75, 0.95$ from an artificial data set.

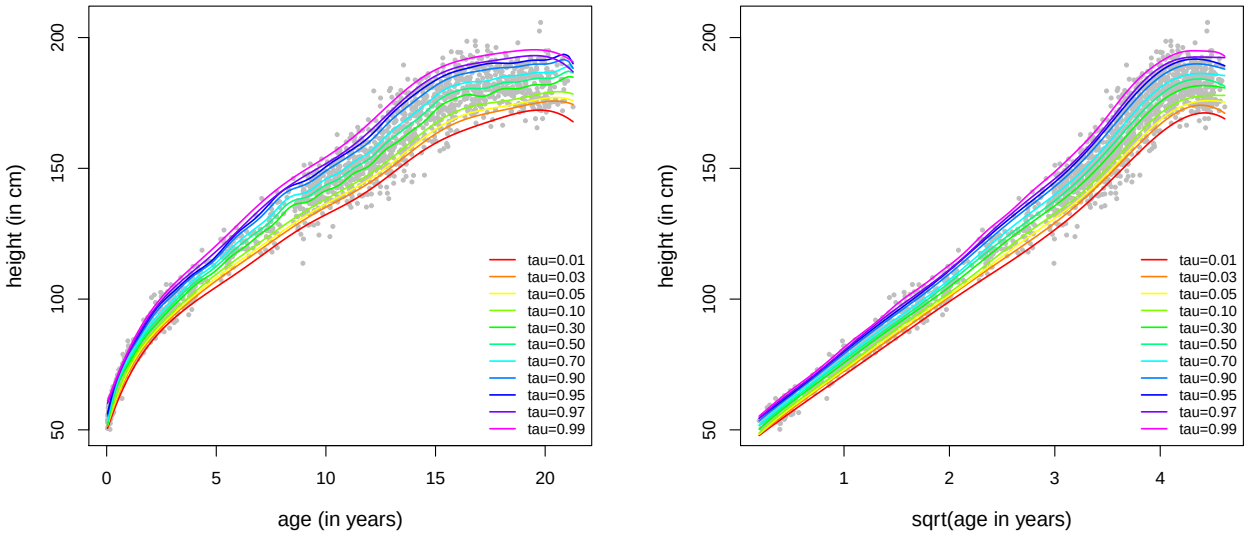

Figure 5.2: Estimated expectiles for $\tau = 0.01, 0.03, 0.05, 0.10, 0.30, 0.50, 0.70, 0.90, 0.95, 0.97, 0.99$ for height against age of HEAD-CIRCUM. The graphs comprise of expectile curves for original data set (left) and data set with transformed age (right).

on the data set size, between 2 and 10 times faster than the solver for quantile regression. Here we note, that since both solvers are part of the same software package mentioned above, the run time comparison is very fair. Indeed both solvers were implemented with the same amount of care and rely on the same framework for e.g. computing the kernel matrices and validation errors. Consequently, if quantile or expectile regression is used as a tool to know the conditional distribution, then kernel-based expectile regression has a clear computational advantage

over quantile regression. In this respect, we like to point to Figure 5.1, in which conditional distributions of an artificial dataset are described by expectiles and quantiles, respectively. We clearly see that the true expectiles do not coincide with the true quantiles, but both the collection of conditional expectiles and of conditional quantiles provide a good overview of the conditional distributions. In addition, both solvers estimate the true expectiles or quantiles with high accuracy, so that at least in this example the computational advantage of expectile regression does not come for the price of reduced accuracy. Finally, as a further illustration, Figure 5.2 presents the expectile curves for different τ considering height against age from data set HEAD-CIRCUM. On the left we see some crossing and wiggling problems. Following Schnabel and Eilers (2009b), the use of square root transformation on age resolves these issues as the right figure shows. This indicates expectiles may be considered as an alternative method to construct different well-known growth charts.

DATA	$\tau = 0.25$			$\tau = 0.50$			$\tau = 0.75$		
	EX-SVM	EX-SVM*	ER-BOOST	EX-SVM	EX-SVM*	ER-BOOST	EX-SVM	EX-SVM*	ER-BOOST
NC-CRIME	0.00616	**0.00555**	0.00948	0.00669	**0.00605**	0.01367	0.00536	**0.00509**	0.01459
	(0.00182)	**(0.00169)**	(0.00177)	(0.00194)	**(0.00161)**	(0.00305)	(0.00172)	**(0.00157)**	(0.00405)
CONCRETE-COMP	0.00901	**0.00893**	0.03961	0.01021	**0.01013**	0.05038	0.00889	**0.00879**	0.04556
	(0.00130)	**(0.00128)**	(0.00365)	(0.00122)	**(0.00117)**	(0.00417)	(0.00112)	**(0.00101)**	(0.00339)
AIRFOIL-NOISE	0.00814	**0.00806**	0.04223	0.00947	**0.00939**	0.04817	0.00855	**0.00850**	0.03832
	(0.00121)	**(0.00119)**	(0.00211)	(0.00134)	**(0.00115)**	(0.00256)	(0.00092)	**(0.00087)**	(0.00218)
MUNICH-RENT	0.00131	**0.00126**	0.01569	0.00122	**0.00121**	0.01812	0.00101	**0.00101**	0.01598
	(0.00033)	**(0.00030)**	(0.00087)	(0.00029)	**(0.00029)**	(0.00113)	(0.00018)	**(0.00016)**	(0.00103)
UPDRS-MOTOR	0.02518	**0.02502**	0.05345	0.02844	**0.02828**	0.06257	0.02585	**0.02569**	0.015229
	(0.00152)	**(0.00152)**	(0.00069)	(0.00159)	**(0.00152)**	(0.00150)	**(0.00166)**	(0.00169)	(0.00179)
HEAD-CIRCUM	0.00323	0.00323	0.02419	0.00390	0.00390	0.02482	**0.00333**	0.00333	0.01855
	(0.00008)	(0.00008)	(0.00047)	(0.00011)	(0.00011)	(0.00057)	**(0.00009)**	(0.00096)	(0.00045)
CYCLE-PP	0.00420	**0.00421**	0.03588	0.00516	0.00516	0.04536	0.00479	**0.00477**	0.03930
	(0.00009)	**(0.00011)**	(0.00079)	(0.00019)	(0.00019)	(0.00097)	**(0.00027)**	(0.00029)	(0.00076)
HOUR	0.01575	**0.01543**	0.02888	0.01664	**0.01627**	0.04021	0.01285	**0.01259**	0.03821
	(0.00029)	(0.00034)	(0.00077)	(0.00046)	**(0.00043)**	(0.00110)	**(0.00031)**	(0.00035)	(0.00103)
CAL-HOUSING	0.02426	**0.02415**	0.05406	0.02546	**0.02518**	0.07473	0.01919	**0.01912**	0.07337
	(0.00126)	**(0.00117)**	(0.00135)	(0.00123)	**(0.00119)**	(0.00158)	(0.00071)	**(0.00064)**	(0.00144)

Table 5.4: Average test error (standard deviation) for 2D-SVM with unclipped duality gap stopping criterion (EX-SVM), 2D-SVM with clipped duality gap stopping criterion (EX-SVM*) and ER-BOOST. The average test error (standard deviation) was computed on 25 random splits for small data sets, 10 random splits for medium size data sets and 5 random splits for larger size data sets.

DATA	τ	EX-SVM	EX-SVM*	Q-SVM	Q-SVM*	ER-BOOST
NC-CRIME	0.25	**0.136**	0.139	0.501	0.530	11.897
	0.50	**0.147**	0.176	0.562	0.599	11.753
	0.75	**0.143**	0.147	0.492	0.535	11.828
CONCRETE-COMP	0.25	**0.354**	0.355	1.394	1.538	1.196
	0.50	**0.357**	0.379	1.430	1.556	1.138
	0.75	**0.339**	0.355	1.353	1.506	1.108
AIRFOIL-NOISE	0.25	0.679	0.759	3.141	3.612	**0.422**
	0.50	0.726	0.779	3.332	3.774	**0.438**
	0.75	0.666	0.703	3.111	3.492	**0.414**
MUNICH-RENT	0.25	**0.559**	0.586	5.878	6.351	6.178
	0.50	**0.577**	0.603	5.338	5.919	6.199
	0.75	**0.606**	0.612	5.398	6.061	6.170
UPDRS-MOTOR	0.25	**13.422**	14.665	34.791	37.971	65.566
	0.50	**15.196**	16.779	36.355	39.290	67.372
	0.75	**13.800**	15.251	33.914	37.158	66.268
HEAD-CIRCUM	0.25	9.226	9.902	57.973	66.068	**0.982**
	0.50	9.874	10.713	49.679	56.460	**0.983**
	0.75	8.862	9.671	58.271	66.785	**0.993**
CYCLE-PP	0.25	20.800	22.516	117.756	131.512	**1.546**
	0.50	22.765	25.365	105.001	116.965	**1.544**
	0.75	20.547	22.603	119.277	133.156	**1.511**
HOUR	0.25	79.635	79.435	375.984	359.189	**48.031**
	0.50	94.416	94.566	328.593	318.889	**47.559**
	0.75	92.486	94.019	317.514	337.308	**48.138**
CAL-HOUSING	0.25	142.244	147.692	614.082	652.405	**22.527**
	0.50	169.433	174.212	527.763	542.678	**22.551**
	0.75	162.605	168.445	549.389	546.775	**21.939**

Table 5.5: Training time (in seconds) for 2D-SVM with unclipped duality gap stopping criterion for expectile (EX-SVM) and quantile (Q-SVM), 2D-SVM with clipped duality gap stopping criterion for expectile (EX-SVM*) and quantile (Q-SVM*), and ER-BOOST.

Appendix A

In Chapter 5 we have discussed some experimental findings achieved by implementing the proposed solver for the expectile regression under different initialization strategies, working set selection strategies, clipping options, and with different numbers of nearest neighbors. This appendix gives the pictorial representation of the performance of the solver under different choices of aforementioned parameters. This helps to see the impact of these choices by comparing the total and (average) per-grid training time (in seconds) and total and (average) per-grid number of iterations. This further leads us to choose the best combination of these parameters that we finally use to compare our solver to the existing package ER-Boost for the expectile regression (see, Chapter 5) in terms of training time and test error and also to the similarly implemented SVM solver for quantile regression in terms of training time.

A.1 Results for Different Working Set Selection Methods

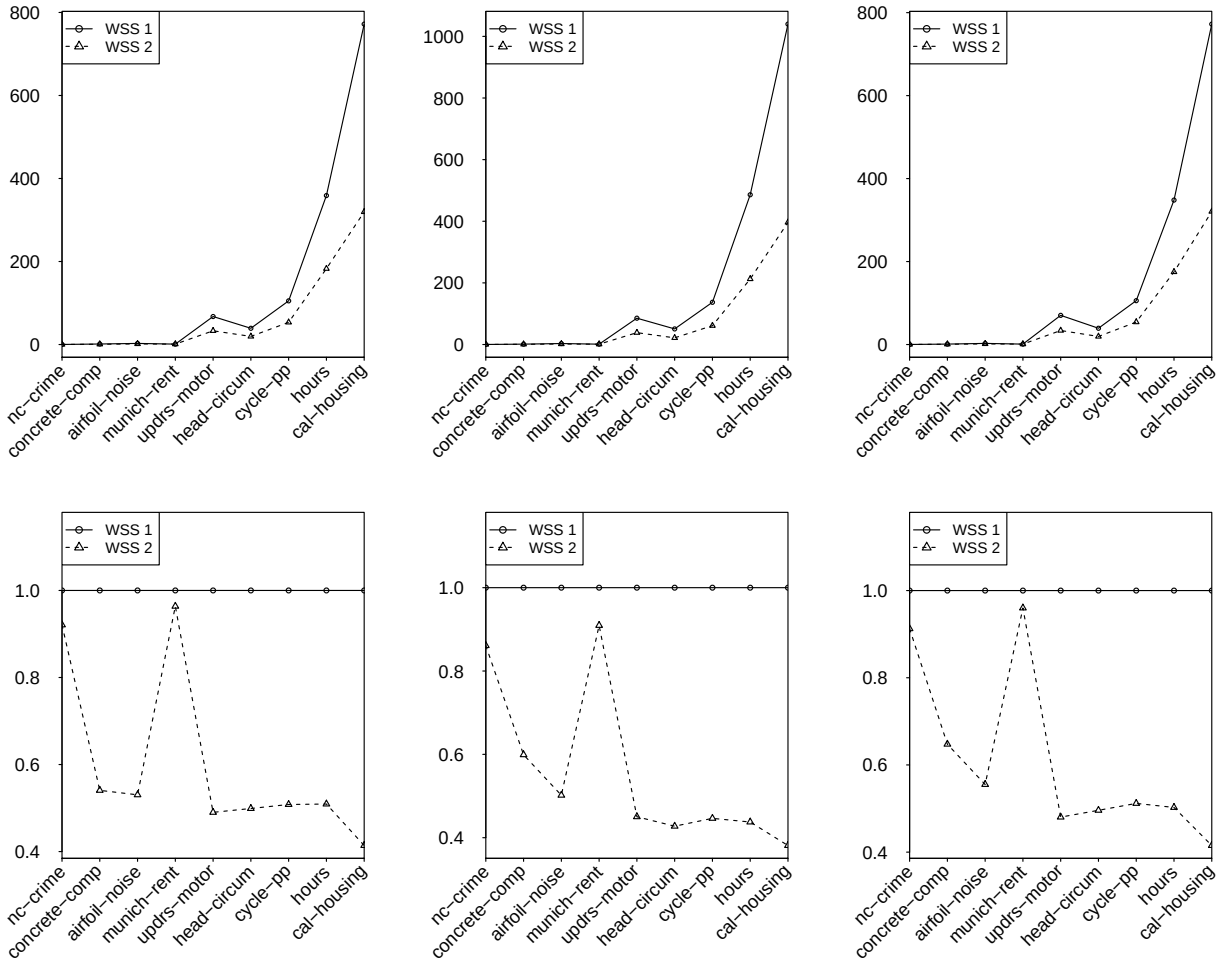

Figure A.1: Training time (top) in seconds and corresponding ratios (bottom) for the two working set selection methods on different data sets. In these experiments, the warm start initialization was chosen and the stopping criterion was based on the clipped duality gap. The graphs show the results for $\tau = 0.25$ (left), $\tau = 0.50$ (middle) and $\tau = 0.75$ (right).

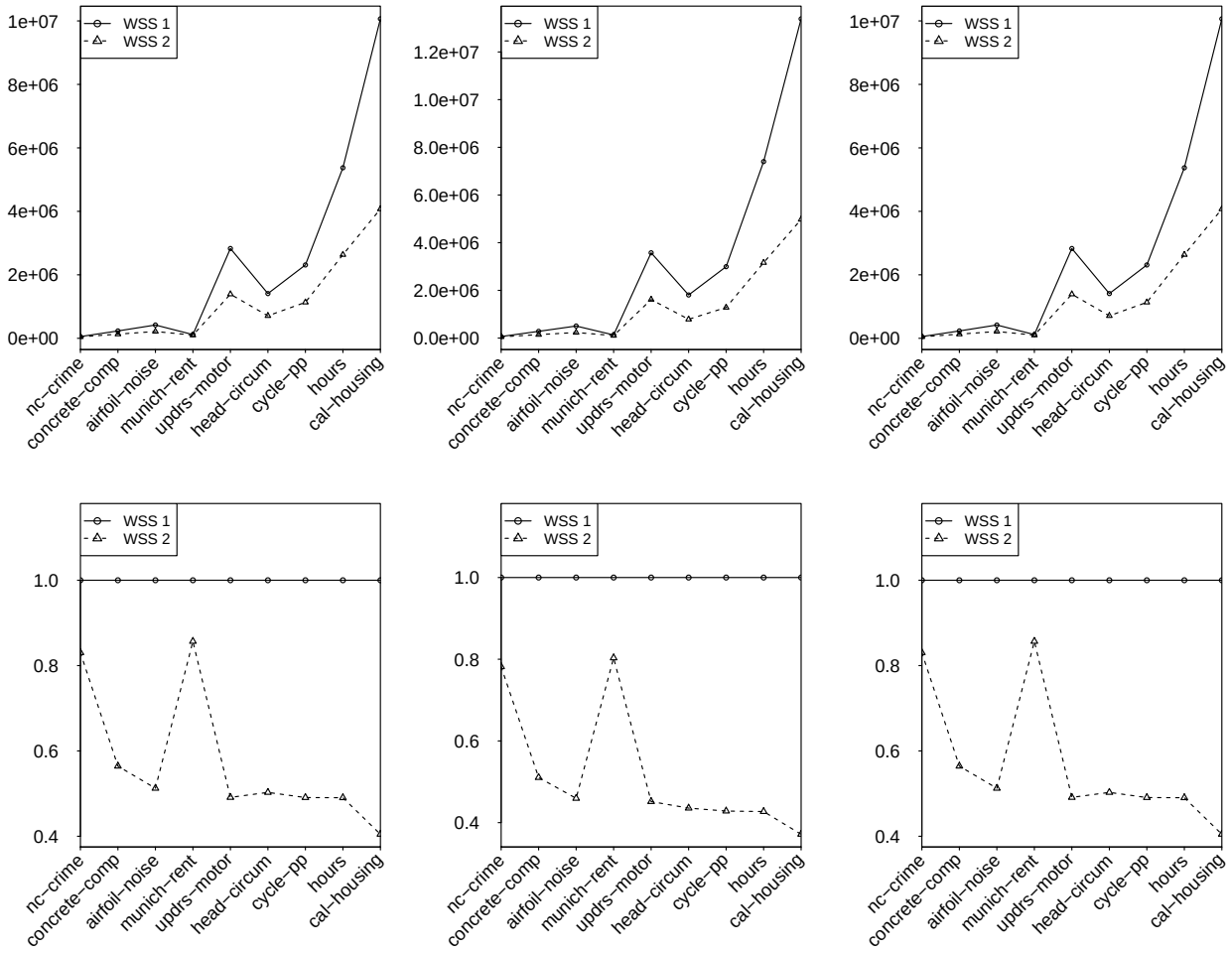

Figure A.2: Training iterations (top) and corresponding ratios (bottom) for the two working set selection methods on different data sets. In these experiments, the warm start initialization was chosen and the stopping criterion was based on the clipped duality gap. The graphs show the results for $\tau = 0.25$ (left), $\tau = 0.50$ (middle) and $\tau = 0.75$ (right).

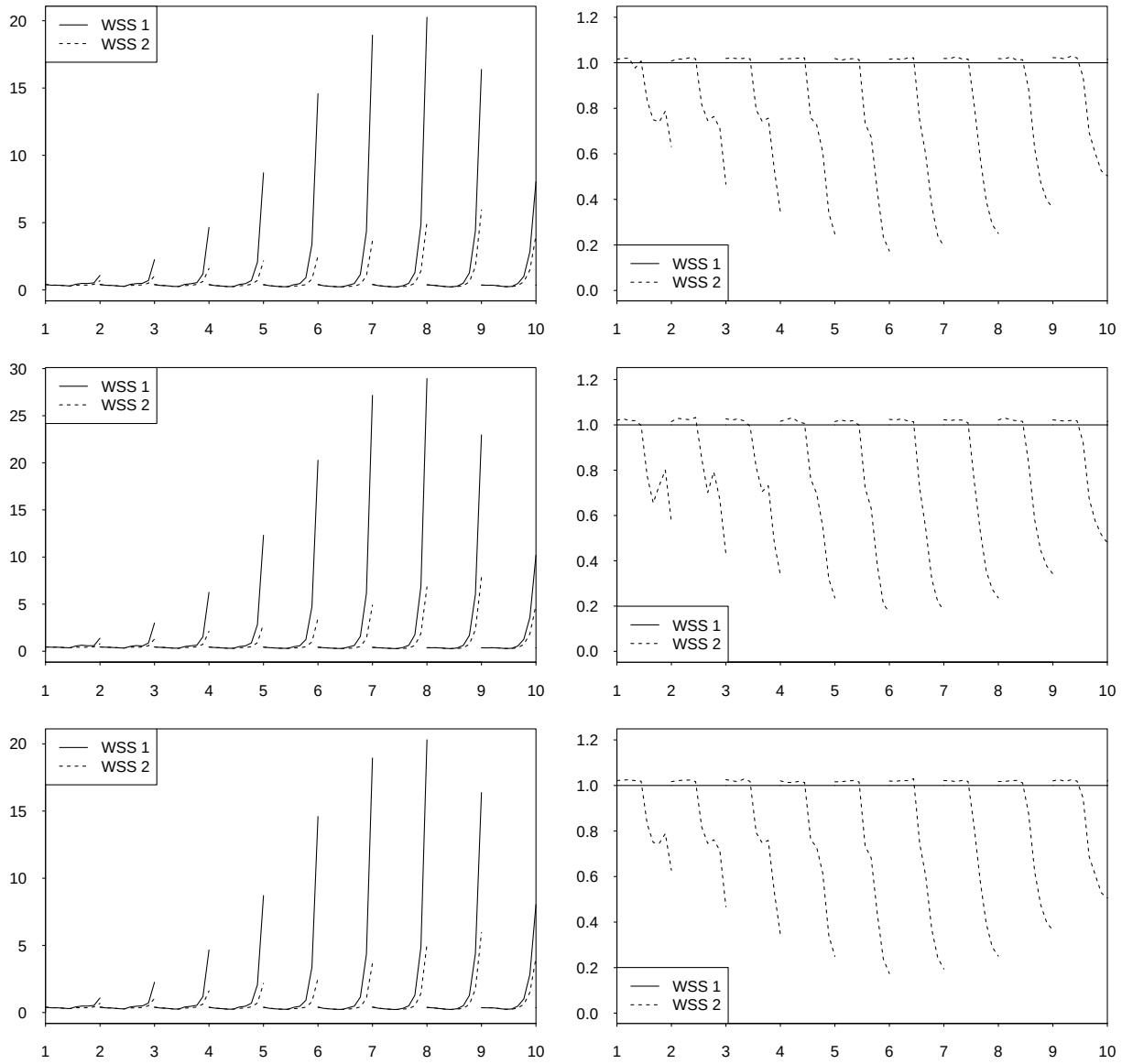

Figure A.3: Average training time in seconds (left) and corresponding ratios (right) per grid point for the two working set selection strategies using the clipped duality gap criterion and the warm start for the CAL-HOUSING data set. For WSS 2, 15 nearest neighbors are considered. The graphs show the results for $\tau = 0.25$ (top), $\tau = 0.50$ (middle) and $\tau = 0.75$ (bottom).

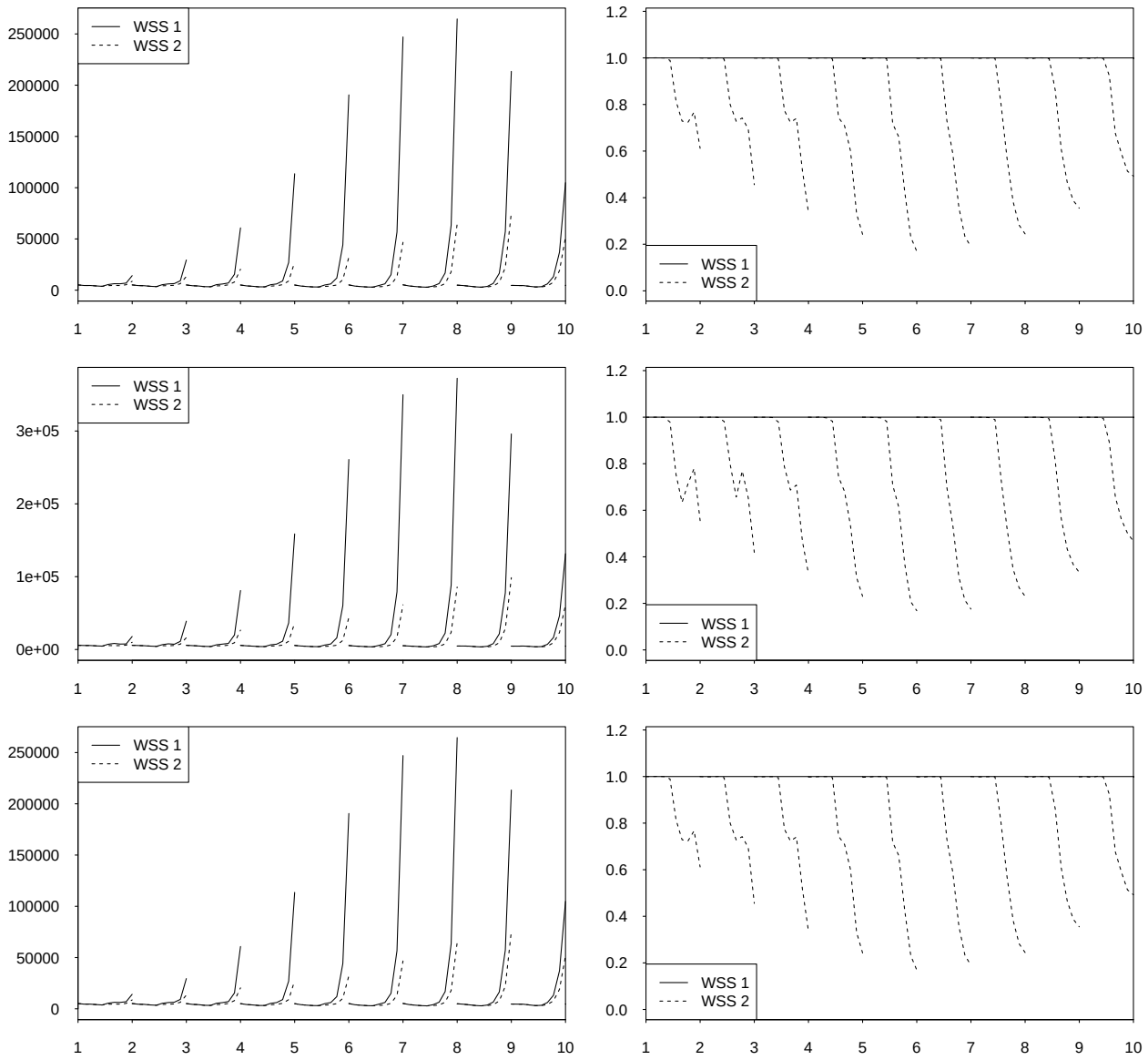

Figure A.4: Average number of iterations (left) and corresponding ratios (right) per grid point for two working set selection strategies using the clipped duality gap criterion and the warm start for the CAL-HOUSING data set. For WSS 2, 15 nearest neighbors are considered. The graphs show the results for $\tau = 0.25$ (top), $\tau = 0.50$ (middle) and $\tau = 0.75$(bottom).

A.2 Results for Different Number of Nearest Neighbors

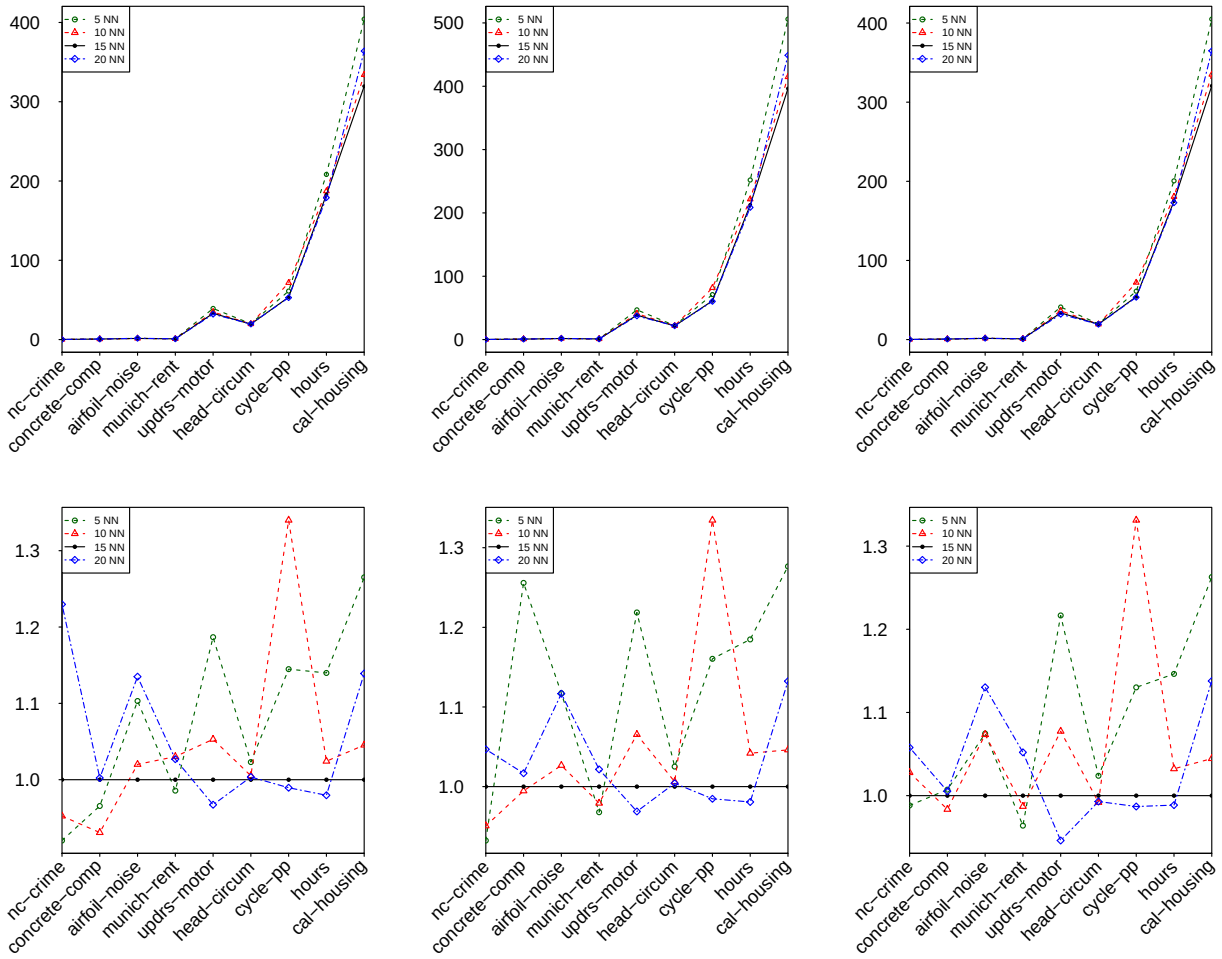

Figure A.5: Training time in seconds (top) and corresponding ratios (bottom) for different numbers of nearest neighbors on different data sets. In these experiments, the warm start initialization was chosen and the stopping criterion was based on the clipped duality gap. The ../plots/chapter5 show the results for $\tau = 0.25$ (left), $\tau = 0.50$ (middle) and $\tau = 0.75$ (right).

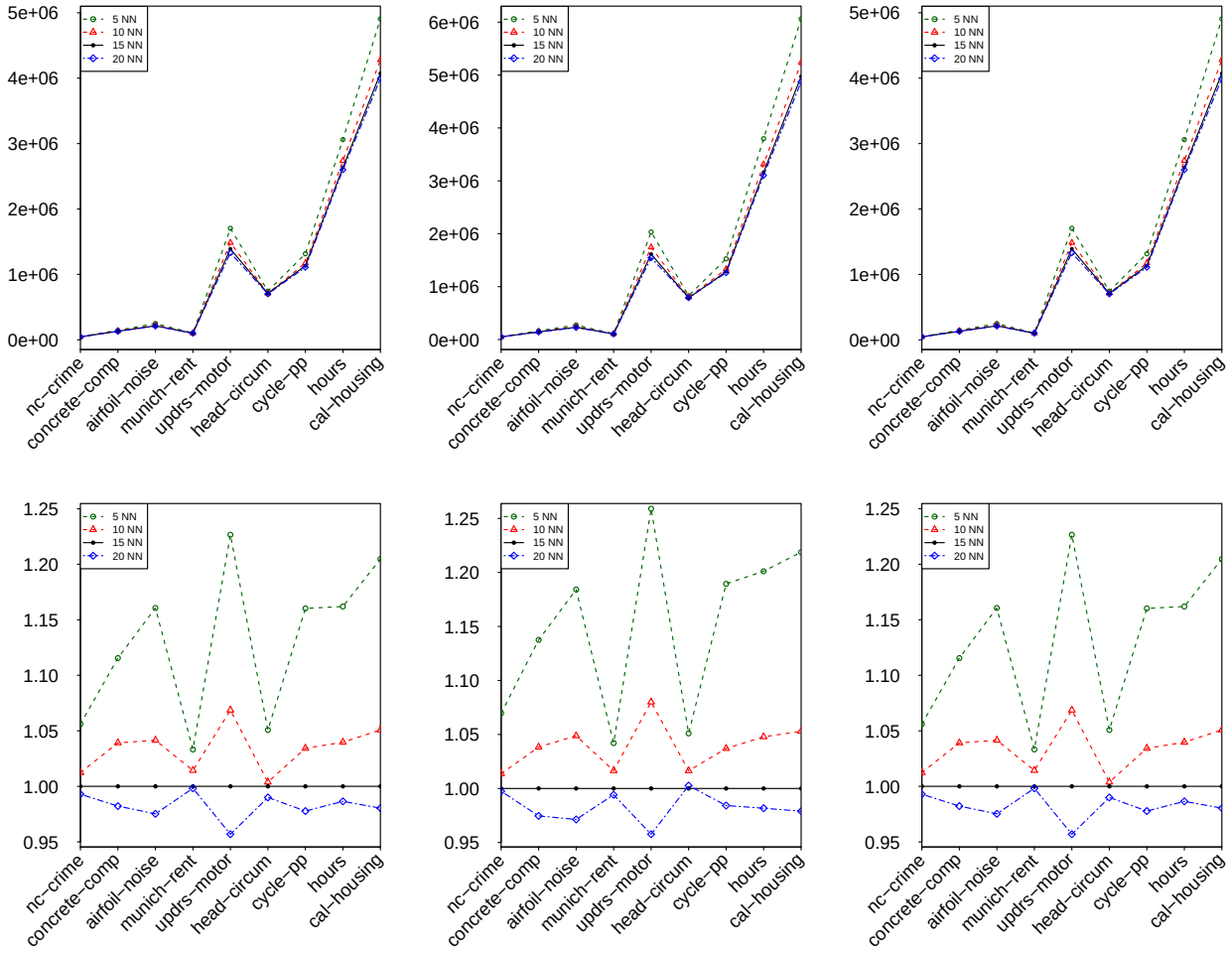

Figure A.6: Training iterations (top) and corresponding ratios (bottom) for different numbers of nearest neighbors on different data sets. In these experiments, the warm start initialization was chosen and the stopping criterion was based on the clipped duality gap. The graphs show the results for $\tau = 0.25$ (left), $\tau = 0.50$ (middle) and $\tau = 0.75$ (right).

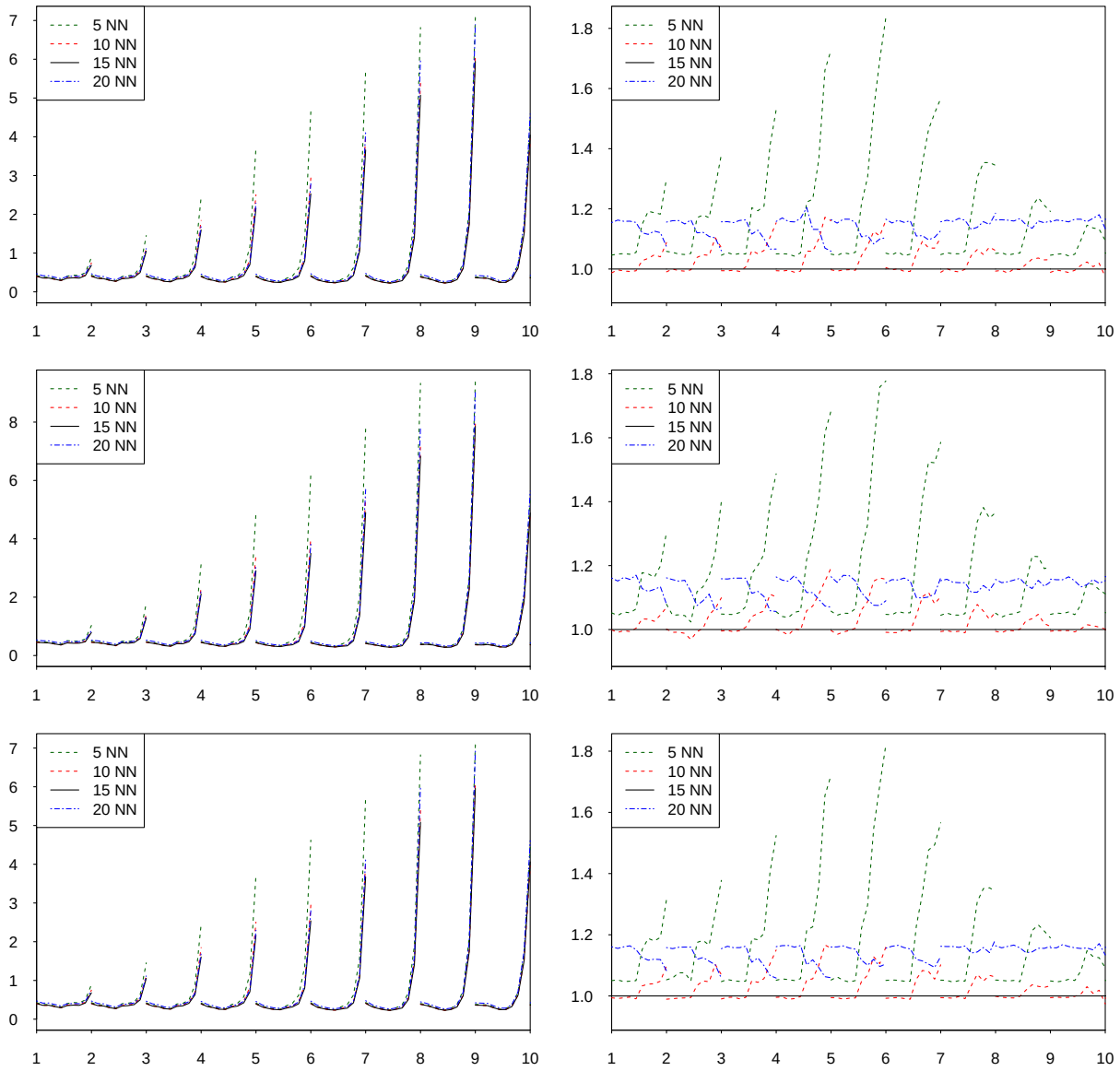

Figure A.7: Average training time in seconds (left) and corresponding ratios (right) per grid point for different numbers of nearest neighbors on the CAL-HOUSING data set. In these experiments, the warm start initialization was chosen and the stopping criterion was based on the clipped duality gap. The graphs show the results for $\tau = 0.25$ (top), $\tau = 0.50$ (middle) and $\tau = 0.75$ (bottom).

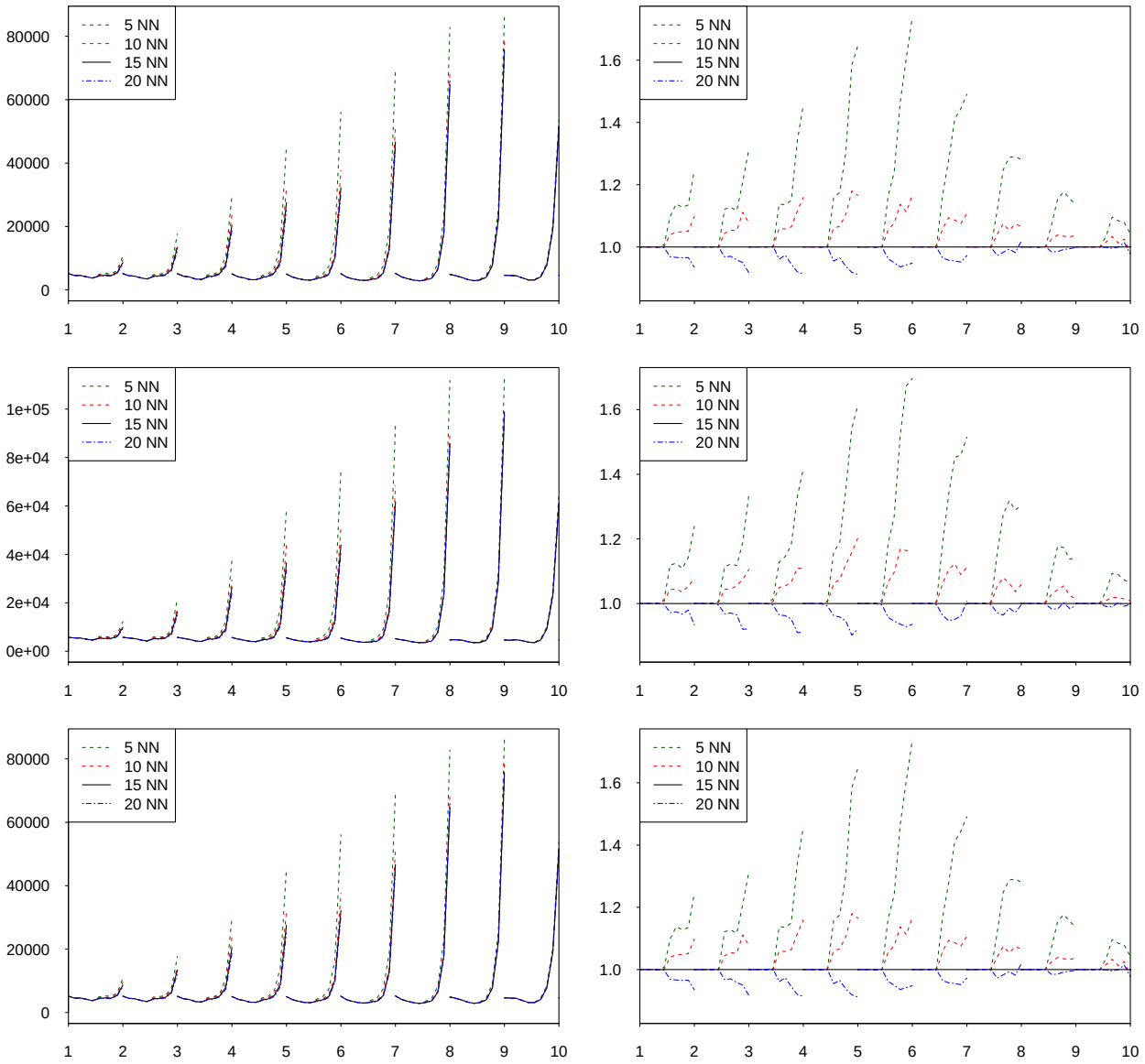

Figure A.8: Average number of iterations (left) and corresponding ratios (right) per grid point for different numbers of nearest neighbors on the CAL-HOUSING data set. In these experiments, the warm start initialization was chosen and the stopping criterion was based on the clipped duality gap. The graphs show the results for $\tau = 0.25$ (top), $\tau = 0.50$ (middle) and $\tau = 0.75$ (bottom).

A.3 Results for Different Initialization Methods

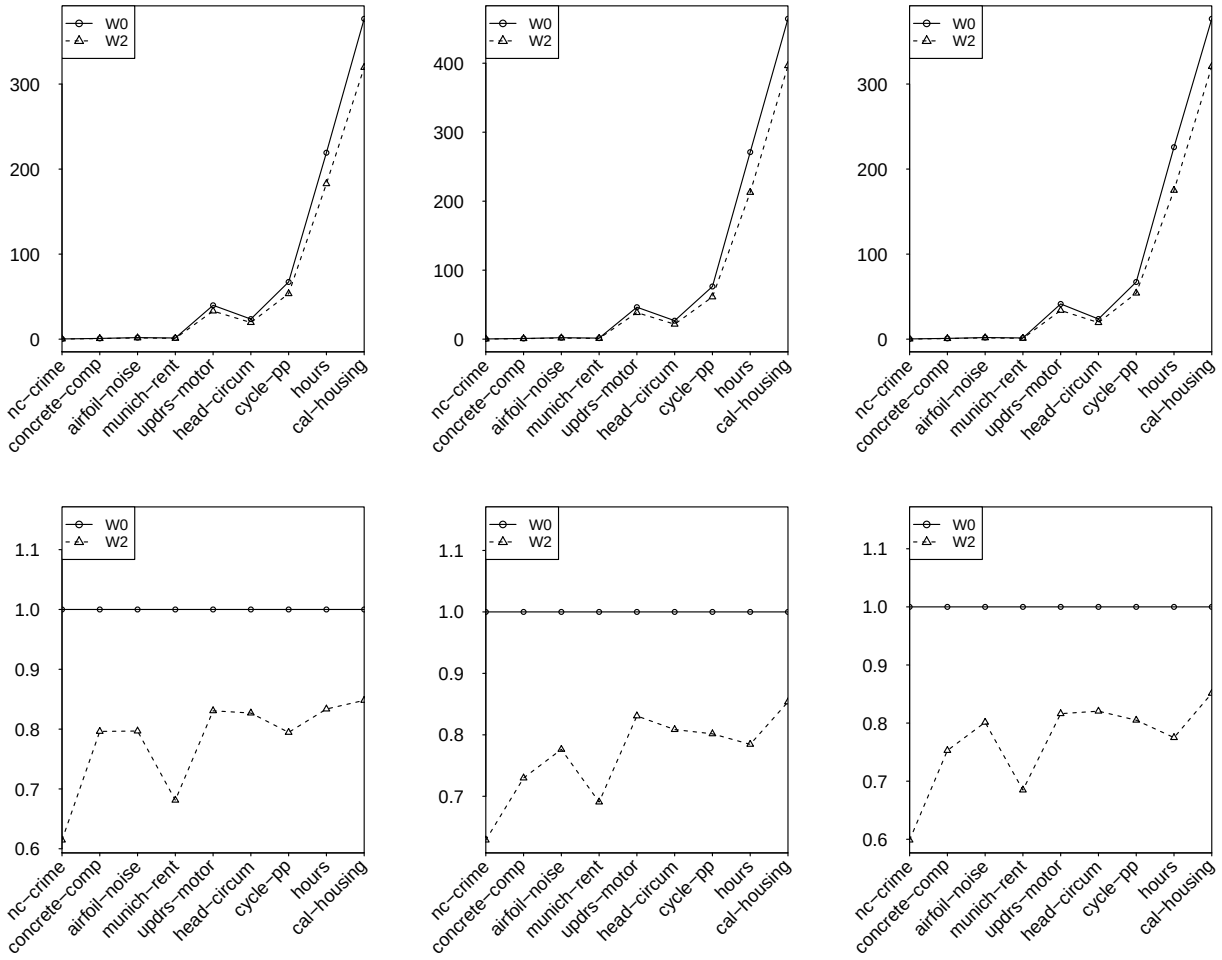

Figure A.9: Training time in seconds (top) and corresponding ratios (bottom) for different initialization methods on different data sets. In these experiments the stopping criterion with clipped duality gap and $NN = 15$ were chosen. The graphs show the results for $\tau = 0.25$ (left), $\tau = 0.50$ (middle) and $\tau = 0.75$ (right).

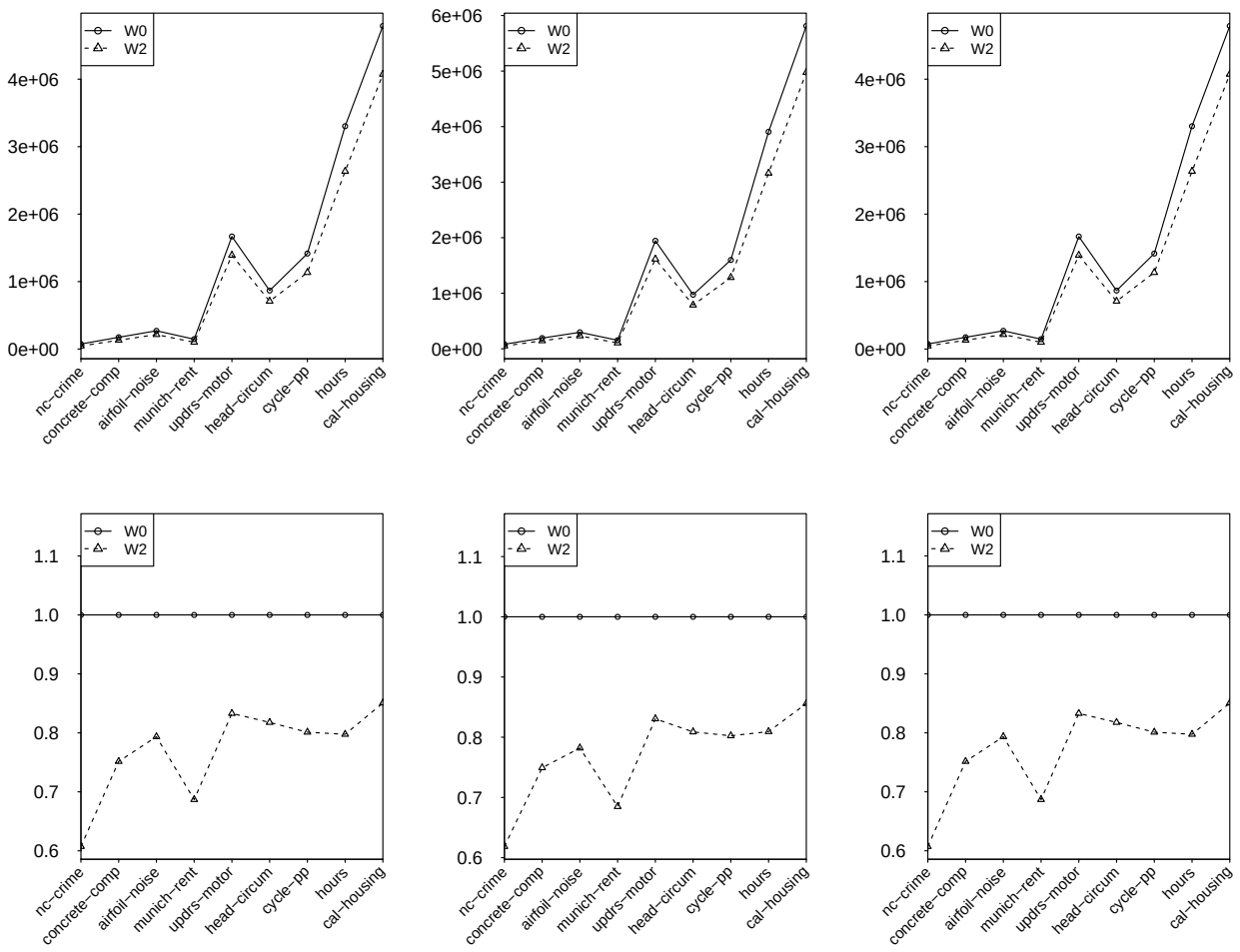

Figure A.10: Training iterations (top) and corresponding ratios (bottom) for different initialization methods on different data sets. In these experiments the stopping criterion with clipped duality gap and WSS 2 with $NN = 15$ were chosen. The graphs show the results for $\tau = 0.25$ (left), $\tau = 0.50$ (middle) and $\tau = 0.75$ (right).

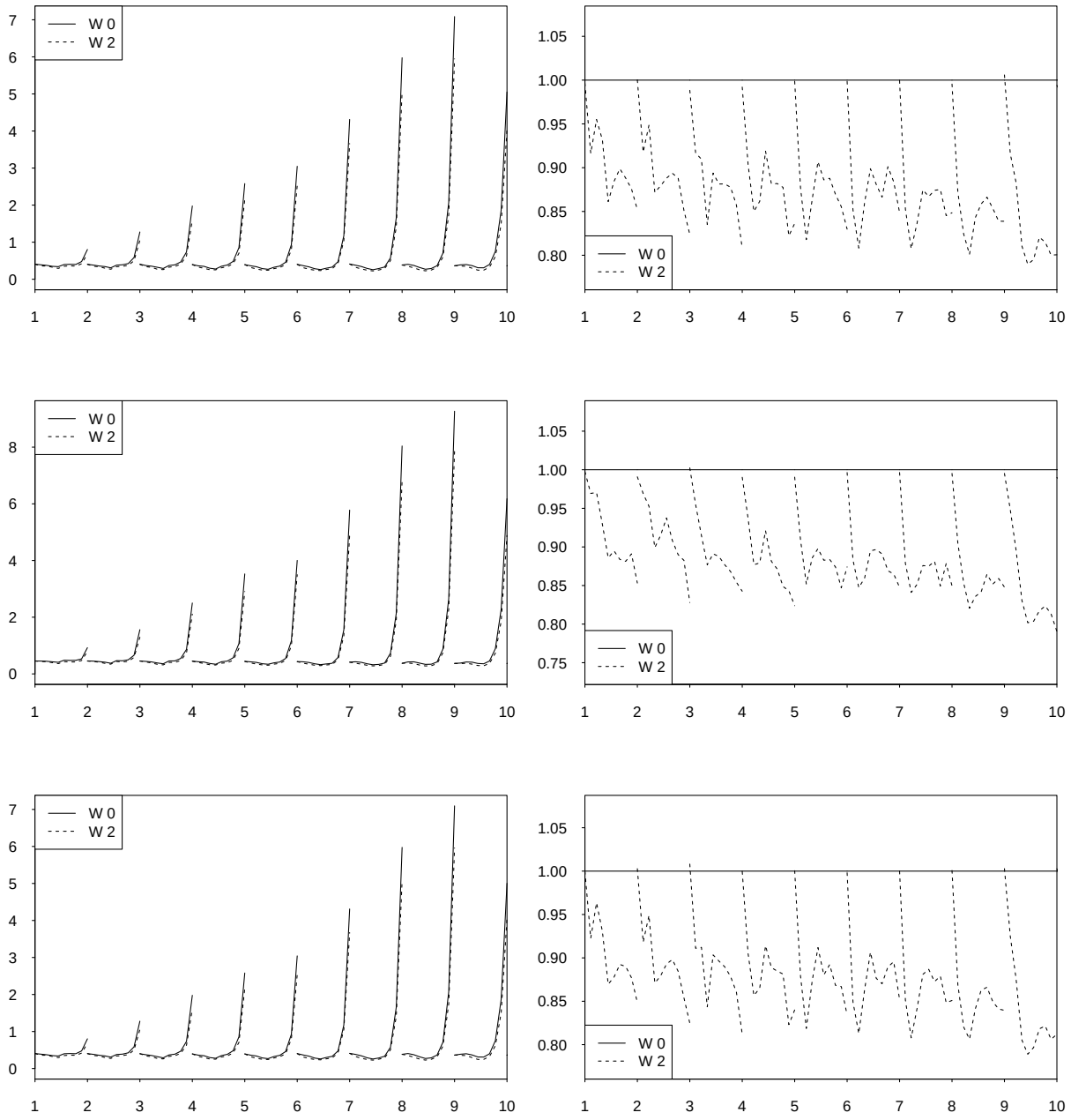

Figure A.11: Average training time in seconds (left) and corresponding ratios (right) per grid point for different initialization methods for the CAL-HOUSING data set. In these experiments the stopping criterion with clipped duality gap and WSS 2 with $NN = 15$ were chosen. The graphs show the results for $\tau = 0.25$ (top), $\tau = 0.50$ (middle) and $\tau = 0.75$ (bottom).

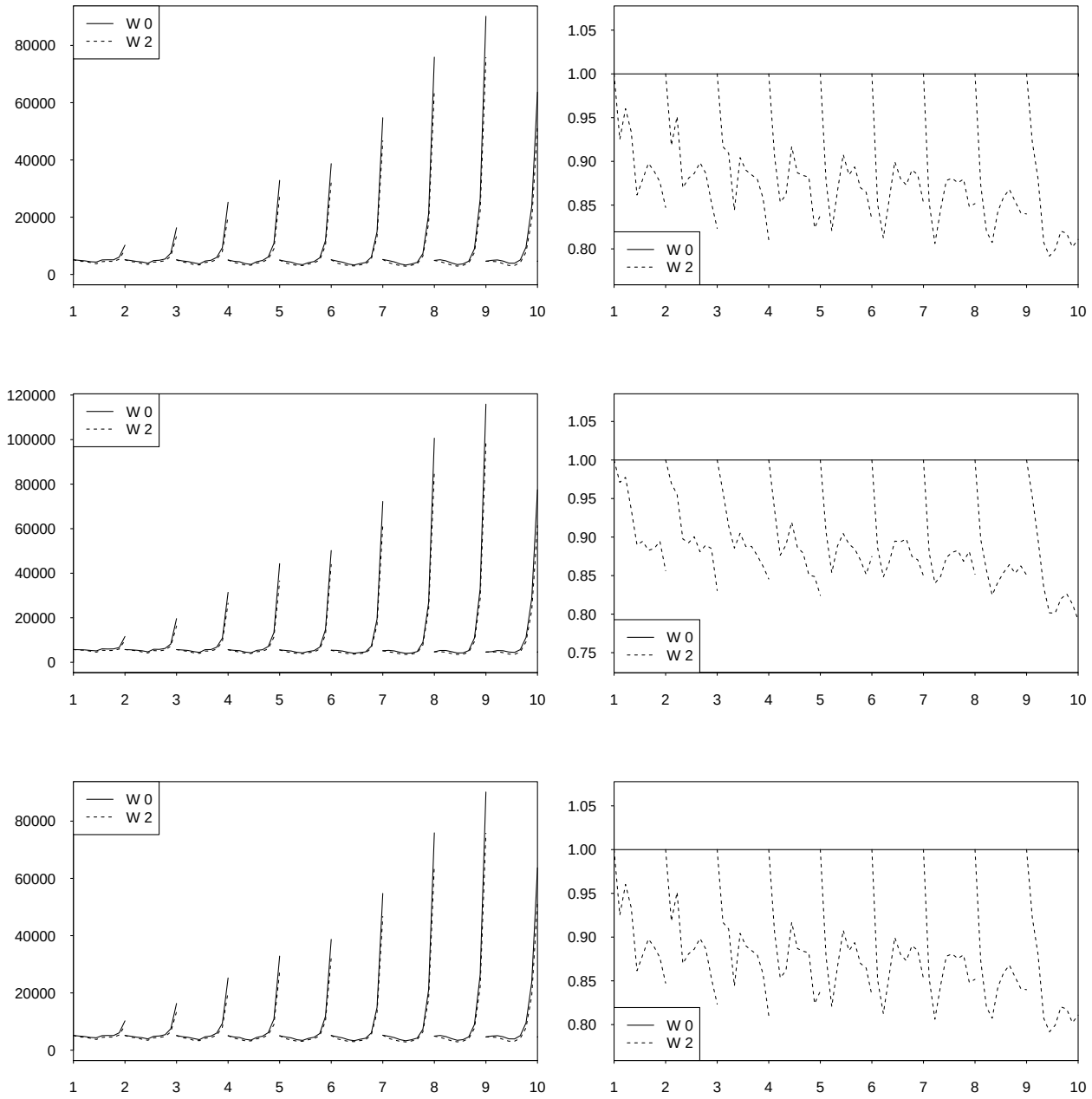

Figure A.12: Average train iterations (left) and corresponding ratios (right) per grid point for different initialization methods for the CAL-HOUSING data set. In these experiments the stopping criterion with clipped duality gap and WSS 2 with $NN = 15$ were chosen. The graphs show the results for $\tau = 0.25$ (top), $\tau = 0.50$ (middle) and $\tau = 0.75$ (bottom).

A.4 Results for Different Stopping Criteria

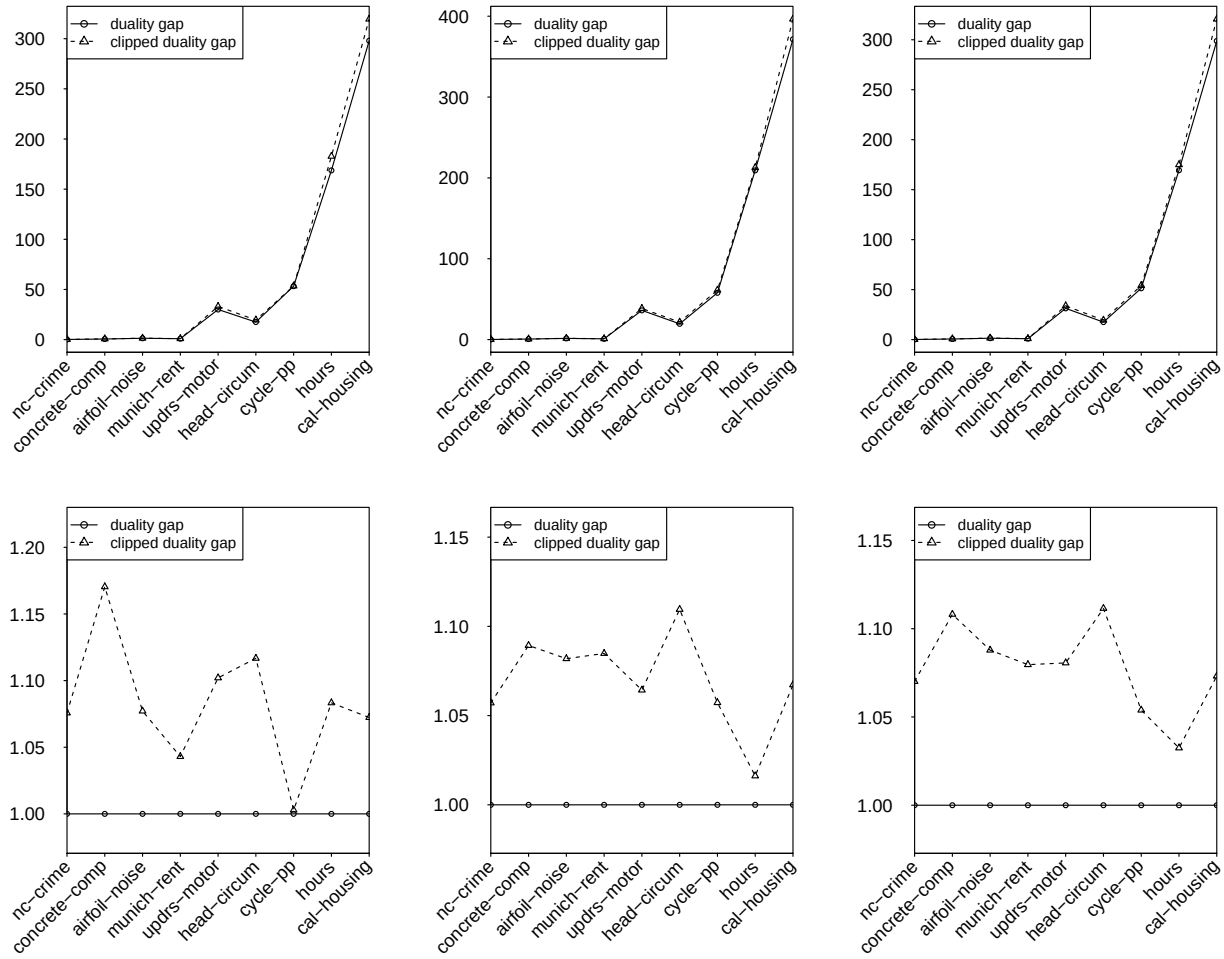

Figure A.13: Training time in seconds (top) and corresponding ratios (bottom) for the two stopping criteria on different data sets. In these experiments the warm start initialization and WSS 2 with $NN = 15$ were chosen. The graphs show the results for $\tau = 0.25$ (left), $\tau = 0.50$ (middle) and $\tau = 0.75$ (right).

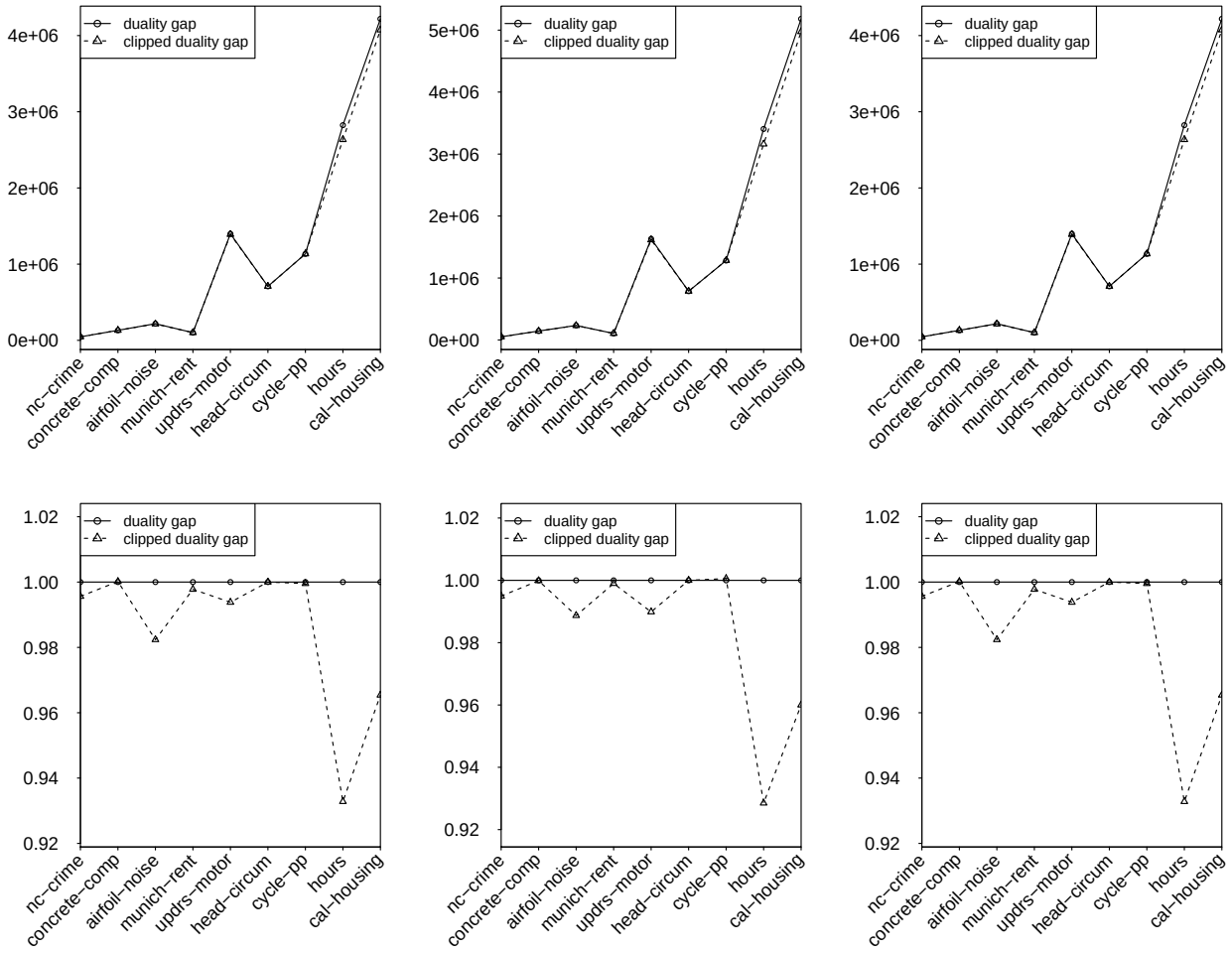

Figure A.14: Training iterations (top) and corresponding ratios (bottom) for the two stopping criteria on different data sets. In these experiments the warm start initialization and WSS 2 with $NN = 15$ were chosen. The graphs show the results for $\tau = 0.25$ (left), $\tau = 0.50$ (middle) and $\tau = 0.75$ (right).

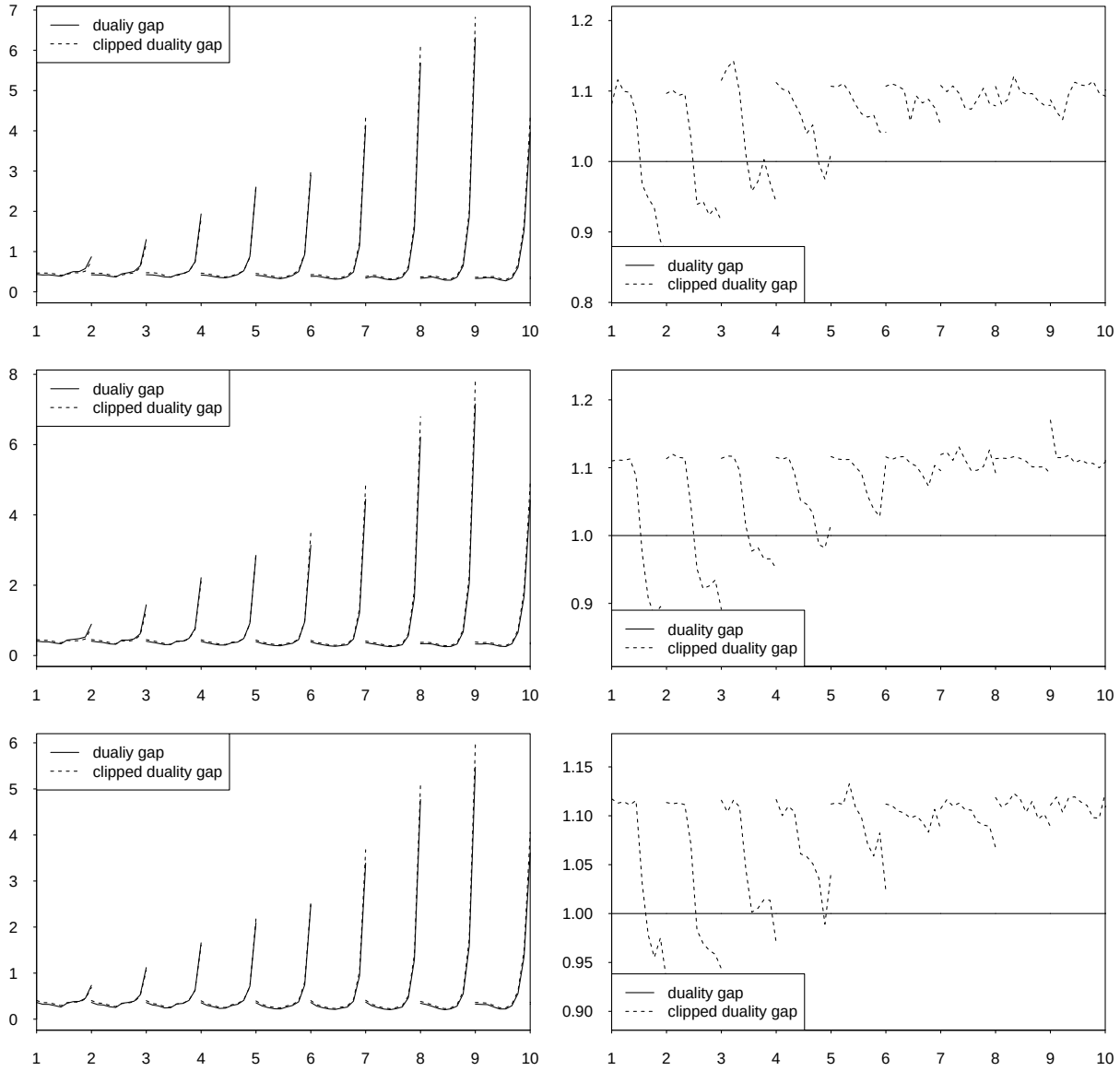

Figure A.15: Average training time in seconds (left) and corresponding ratios (right) per grid point for the two stopping criterion on the CAL-HOUSING data set. In these experiments the warm start initialization and WSS 2 with $NN = 15$ were chosen. The graphs show the results for $\tau = 0.25$ (top), $\tau = 0.50$ (middle) and $\tau = 0.75$ (bottom).

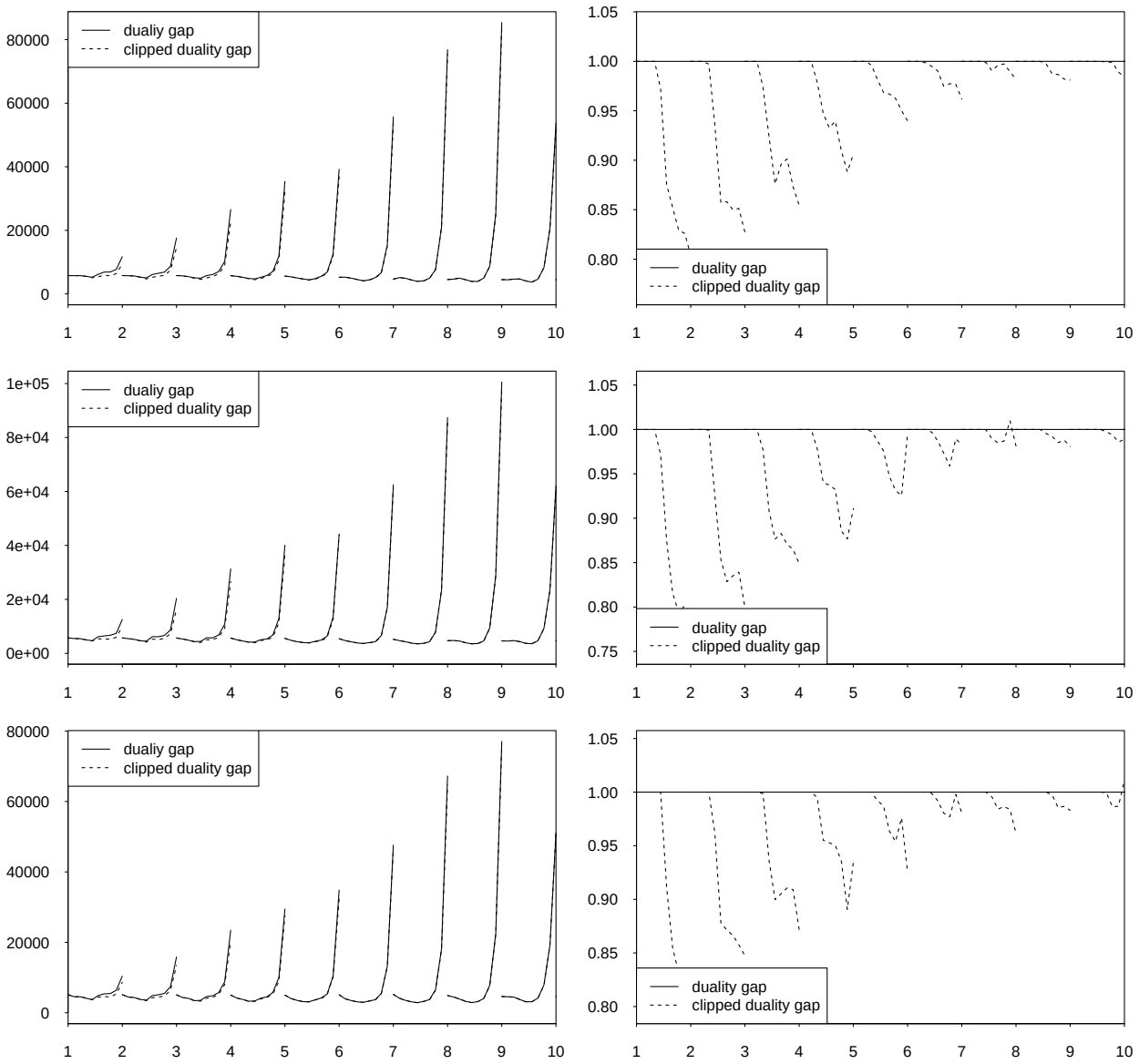

Figure A.16: Average number of training iterations (left) and corresponding ratios (right) per grid point for the two stopping criterion on the CAL-HOUSING data set. In these experiments the warm start initialization and WSS 2 with $NN = 15$ were chosen. The graphs show the results for $\tau = 0.25$ (top), $\tau = 0.50$ (middle) and $\tau = 0.75$ (bottom).

Bibliography

Abdous B, Remillard B (1995) Relating quantiles and expectiles under weighted-symmetry. Ann Inst Statist Math 47:371–384, http://dx.doi.org/10.1007/bf00773468

Abe S (2005) Support Vector Machines for Pattern Classification, vol 2. Springer, https://doi.org/10.1007/978-1-84996-098-4

Adams RA, Fournier JJF (2003) Sobolev Spaces. Academic Press, New York, 2nd edition, https://doi.org/10.1016/s0079-8169(03)x8001-0

Aragon Y, Casanova S, Chambers R, Leconte E (2005) Conditional ordering using nonparametric expectiles. J Off Stat 21:617–633, http://www.jos.nu/Articles/abstract.asp?article=214617

Aronszajn N (1950) Theory of reproducing kernels. Trans Amer Math Soc 68:337–404

Bauer F, Pereverzev S, Rosasco L (2007) On regularization algorithms in learning theory. J complexity 23:52–72, https://doi.org/10.1016/j.jco.2006.07.001

Bellini F, Klar B, Müller A, Gianin RE (2014) Generalized quantiles as risk measures. Insurance Math Econom 54:41–48, http://dx.doi.org/10.1016/j.insmatheco.2013.10.015

Bennett C, Sharpley RC (1988) Interpolation of Operators. Academic Press, Boston, https://doi.org/10.1016/s0079-8169(08)x6053-2

Blanchard G, Bousquet O, Massart P (2008) Statistical performance of support vector machines. Ann Statist 36:489–531, https://doi.org/10.1214/009053607000000839

Bousquet O, Elisseeff A (2002) Stability and generalization. J Mach Lear Res 2:499–526

Boyd S, Vandenberghe L (2004) Convex Optimization. Cambridge university press, https://doi.org/10.1017/cbo9780511804441

Breckling J, Chambers R (1988) M-quantiles. Biometrika 75:761–771, http://dx.doi.org/10.2307/2336317

Caponnetto A, De Vito E (2007) Optimal rates for the regularized least-squares algorithm. Found Comput Math 7:331–368, https://doi.org/10.1007/s10208-006-0196-8

Chang CC, Lin CJ (2011) LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol 2:27, http://dx.doi.org/10.1145/1961189.1961199

Chen D, Wu Q, Ying Y, Zhou D (2004) Support vector machine soft margin classifiers: error analysis. J Mach Learn Res 5:1143–1175

Christmann A, Steinwart I (2007) How SVMs can estimate quantiles and the median. In: Advances in neural information processing systems, pp 305–312

Cristianini N, Shawe-Taylor J (2000) An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge University Press, Cambridge, http://dx.doi.org/10.1017/cbo9780511801389

Cucker F, Smale S (2002) On the mathematical foundations of learning. Bull Amer Math Soc 39:1–49, https://doi.org/10.1090/s0273-0979-01-00923-5

Cucker F, Zhou DX (2007) Learning Theory: An Approximation Theory Viewpoint. Cambridge University Press, Cambridge, https://doi.org/10.1017/cbo9780511618796

De Vito E, Caponnetto A, Rosasco L (2005) Model selection for regularized least-squares algorithm in learning theory. Foun Comput Math 5:59–85, https://doi.org/10.1007/s10208-004-0134-1

DeVore RA (1998) Nonlinear approximation. Acta numerica 7:51–150,

DeVore RA, Popov VA (1988) Interpolation of besov spaces 305:397–414

DeVore RA, Sharpley RC (1993) Besov spaces on domains in $\mathbb{R}^d$. Trans Amer Math Soc 335:843–864

Devroye L, Györfi L, Lugosi G (1996) A Probabilistic Theory of Pattern Recognition, vol 31. https://doi.org/10.1007/978-1-4612-0711-5

Eberts M, Steinwart I (2013) Optimal regression rates for SVMs using gaussian kernels. Electron J Stat 7:1–42, http://dx.doi.org/10.1214/12-ejs760

Edmunds DE, Triebel H (2008) Function Spaces, Entropy Numbers, Differential Operators. Cambridge University Press, Cambridge, https://doi.org/10.1017/cbo9780511662201

Efron B (1991) Regression percentiles using asymmetric squared error loss. Statist Sci 1:93–125

Farooq M, Steinwart I (2017a) Learning rates for kernel-based expectile regression https://arxiv.org/abs/1702.07552

Farooq M, Steinwart I (2017b) An SVM-like approach for expectile regression. Comput Stat Data Anal 109:159–181, https://doi.org/10.1016/j.csda.2016.11.010

Gneiting T (2011) Making and evaluating point forecasts. J Am Stat Assoc 106:746–762, http://dx.doi.org/10.1198/jasa.2011.r10138

Guler K, Ng PT, Xiao Z (2014) Mincer-Zarnovitz quantile and expectile regressions for forecast evaluations under asymmetric loss functions. Northern Arizona University, The WA Franke College of Business Working Paper Series 14-01

Györfi L, Kohler M, Krzyzak A, Walk H (2002) A Distribution-Free Theory of Nonparametric Regression. Springer-Verlag, New York

Hamidi B, Maillet B, Prigent JL (2014) A dynamic autoregressive expectile for time-invariant portfolio protection strategies. J Econom Dynam Control 46:1–29, http://dx.doi.org/10.1016/j.jedc.2014.05.005

Huang X, Shi L, Suykens JA (2014) Asymmetric least squares support vector machine classifiers. Comput Statist Data Anal 70:395–405, http://dx.doi.org/10.1016/j.csda.2013.09.015

Hush D, Kelly P, Scovel C, Steinwart I (2006) QP algorithms with guaranteed accuracy and run time for support vector machines. J Mach Learn Res 7:733–769

Joachims T (1999) Making large-scale SVM learning practical. In: Advances in Kernel Methods - Support Vector Learning, MIT Press, Cambridge, MA, USA, pp 169–184

Jones MC (1994) Expectiles and M-quantiles are quantiles. Stat Probab Lett 20(2):149–153, https://doi.org/10.1016/0167-7152(94)90031-0

Keating C, Shadwick WF (2002) A universal performance measure. J Perform Meas 6:59–84

Keerthi SS, Chapelle O, DeCoste D (2006) Building support vector machines with reduced classifier complexity. J Mach Learn Res 7:1493–1515

Kim M, Lee S (2016) Nonlinear expectile regression with application to value-at-risk and expected shortfall estimation. Comput Stat Data Anal 94:1–19, https://doi.org/10.1016/j.csda.2015.07.011

Koenker R (1992) When are expectiles percentiles? Econometric Theory 8:423–424

Koenker R (2005) Quantile Regression. Cambridge University Press, Cambridge, http://dx.doi.org/10.1017/cbo9780511754098

Koenker R, Bassett Jr G (1978) Regression quantiles. Econometrica 46:33–50, http://dx.doi.org/10.2307/1913643

Meister M, Steinwart I (2016) Optimal learning rates for localized SVMs. J Mach Learn Res 17:1–44

Mendelson S, Neeman J (2010) Regularization in kernel learning. Ann Statist 38:526–565, https://doi.org/10.1214/09-aos728

Mosteller F, Tukey JW (1977) Data Analysis and Regression: A Second Course in Statistics. Addison-Wesley, Reading, MA,

Newey WK, Powell JL (1987) Asymmetric least squares estimation and testing. Econometrica 55:819–847, http://dx.doi.org/10.2307/1911031

Nikol'skii SM (2012) Approximation of Functions of Several Variables and Imbedding Theorems, vol 205. Springer Science & Business Media, https://doi.org/10.1007/978-3-642-65711-5

Platt J (1999) Fast training of support vector machines using sequential minimal optimization. In: Advances in kernel methods-Support Vector Learning, MIT press, Cambridge, MA., pp 185–208

Schnabel S, Eilers P (2009a) An analysis of life expectancy and economic production using expectile frontier zones. Demographic Res 21:109–134, http://dx.doi.org/10.4054/demres.2009.21.5

Schnabel SK, Eilers PH (2009b) Optimal expectile smoothing. Comput Statist Data Anal 53:4168–4177, http://dx.doi.org/10.1016/j.csda.2009.05.002

Schölkopf B, Smola A (2002) Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT press, Cambridge, MA., https://doi.org/10.1109/tnn.2005.848998

Shawe-Taylor J, Cristianini N (2004) Kernel Methods for Pattern Analysis. Cambridge University Press, Cambridge, https://doi.org/10.1017/cbo9780511809682

Sobotka F, Kneib T (2012) Geoadditive expectile regression. Comput Statist Data Anal 56:755–767, http://dx.doi.org/10.1016/j.csda.2010.11.015

Sobotka F, Radice R, Marra G, Kneib T (2013) Estimating the relationship between women's education and fertility in Botswana by using an instrumental variable approach to semiparametric expectile regression. J Roy Stat Soc C- App 62:25–45, http://dx.doi.org/10.1111/j.1467-9876.2012.01050.x

Stahlschmidt S, Eckardt M, Härdle WK (2014) Expectile treatment effects: An efficient alternative to compute the distribution of treatment effects. Tech. rep., Sonderforschungsbereich 649, Humboldt University, Berlin, Germany

Steinwart I (2003) Sparseness of support vector machines. J Mach Lear Res 4:1071–1105

Steinwart I (2007) How to compare different loss functions and their risks. Constr Approx 26:225–287, https://doi.org/10.1007/s00365-006-0662-3

Steinwart I (2009) Oracle inequalities for support vector machines that are based on random entropy numbers. J Complexity 25:437–454, https://doi.org/10.1016/j.jco.2009.06.002

Steinwart I, Christmann A (2008) Support Vector Machines. Springer, New York, http://dx.doi.org/10.1007/978-0-387-77242-4

Steinwart I, Christmann A (2011) Estimating conditional quantiles with the help of the pinball loss. Bernoulli 17:211–225, http://dx.doi.org/10.3150/10-bej267

Steinwart I, Scovel C (2012) Mercer's theorem on general domains: on the interaction between measures, kernels, and RKHSs. Constr Approx 35:363–417, http://dx.doi.org/10.1007/s00365-012-9153-3

Steinwart I, Thomann P (2017) LiquidSVM: A fast and versatile SVM package, https://arxiv.org/abs/1702.06899

Steinwart I, Hush D, Scovel C (2007) An oracle inequality for clipped regularized risk minimizers. In: Advances in neural information processing systems, pp 1321–1328

Steinwart I, Hush DR, Scovel C (2009) Optimal rates for regularized least squares regression. In: 22nd Annual Conference on Learning Theory, pp 79–93

Steinwart I, Hush D, Scovel C (2011) Training SVMs without offset. J Mach Learn Res 12:141–202

Steinwart I, Pasin C, Williamson R, Zhang S (2014) Elicitation and identification of properties. In: Balcan MF, Szepesvari C (eds) JMLR Workshop and Conference Proceedings Volume 35: Proceedings of the 27th Conference on Learning Theory 2014, pp 482–526

Tacchetti A, Mallapragada PK, Santoro M, Rosasco L (2013) GURLS: a least squares library for supervised learning. J Mach Learn Res 14:3201–3205

Takeuchi I, Le QV, Sears TD, S AJ (2006) Nonparametric quantile estimation. J Mach Learn Res 7:1231–1264

Tartar L (2007) An Introduction to Sobolev Spaces and Interpolation Spaces, vol 3. Springer Science & Business Media, https://doi.org/10.1007/978-3-540-71483-5

Taylor JW (2008) Estimating value at risk and expected shortfall using expectiles. J Financ Econ 6:231–252, http://dx.doi.org/10.1093/jjfinec/nbn001

van der Vaart AW, van Zanten JH (2009) Adaptive Bayesian estimation using a Gaussian random field with inverse Gamma bandwidth. Ann Statist 37:2655–2675, https://doi.org/10.1214/08-aos678

Vapnik V (2000) The Nature of Statistical Learning Theory. Springer-Verlag, New York, http://dx.doi.org/10.1007/978-1-4757-2440-0

Vogt M (2002) SMO algorithms for support vector machines without bias term. Technische Univ Darmstadt, Inst Automat Contr, Lab Contr Syst Process Automat, Darmstadt, Germany

Wang Y, Wang S, Lai KK (2011) Measuring financial risk with generalized asymmetric least squares regression. Appl Soft Comput 11:5793–5800, http://dx.doi.org/10.1016/j.asoc.2011.02.018

Wright S, Nocedal J (1999) Numerical Optimization, vol 2. Springer, New York, http://dx.doi.org/10.1007/b98874

Wu Q, Ying Y, Zhou DX (2006) Learning rates of least-square regularized regression. Found Comput Math 6:171–192, https://doi.org/10.1007/s10208-004-0155-9

Xu Q, Liu X, Jiang C, Yu K (2016) Nonparametric conditional autoregressive expectile model via neural network with applications to estimating financial risk. Appl Stoch Model Bus 32:882–908, https://doi.org/10.1002/asmb.2212

Yang Y, Zou H (2015) Nonparametric multiple expectile regression via ER-Boost. J Stat Comput Simulation 85:1442–1458, http://dx.doi.org/10.1080/00949655.2013.876024

Yao Q, Tong H (1996) Asymmetric least squares regression estimation: a nonparametric approach. J Nonparametr Statist 6:273–292, http://dx.doi.org/10.1080/10485259608832675

Ziegel JF (2016) Coherence and elicitability. Math Financ 26:901–918, https://doi.org/10.1111/mafi.12080

Nomenclature

$\mathbb{R}$	set of real numbers
$\mathbb{N}$	set of natural numbers
$A^c, \mathring{A}, \bar{A}$	complement, interior and closure of a set A
∂A	boundary of a set A, i.e., $\partial A = \bar{A} \setminus \mathring{A}$
X	space of input values, usually a subset of $\mathbb{R}^d$
Y	space of output values, usually a subset of $\mathbb{R}$
d	dimensions of input space
n	sample size
n_{train}, n_{test}	sample size of training and test data set
M	a positive constant
D	data set consisting of the samples $(x_1, y_1), \ldots, (x_n, y_n)$
D	empirical distribution w.r.t. the data set D
$\mathcal{D}$	dual objective function
L_α	asymmetric Least absolute deviation (ALAD) loss function for $\alpha \in (0,1)$
L_τ	asymmetric least square (ALS) loss function for $\tau \in (0,1)$
$L_\tau \circ f$	loss L_τ combined with f, i.e. $L_\tau \circ f(x,y) = L_\tau(y, f(x))$
$\widehat{t}$	clipping operation
$\lvert L_\tau \rvert_{1,M}$	local Lipschitz constant of the loss L_τ for some $M > 0$
$\mathcal{P}$	primal objective function
P, Q	probability distribution
P_X	marginal distribution
$\mathrm{P}(\cdot\lvert x)$	regular conditional distribution
$\mathcal{C}_{L_\tau,Q}(\cdot)$	inner L_τ-risk w.r.t. distribution Q
$\mathcal{C}^\star_{L_\tau,Q}$	minimal inner L_τ-risk w.r.t. distribution Q
$\mathcal{R}_{L_\tau,\mathrm{P}}(\cdot)$	L_τ-risk w.r.t. distribution P
$\mathcal{R}^*_{L_\tau,\mathrm{P}}$	minimal L_τ-risk w.r.t. distribution P
$\mathcal{R}_{L,\mathrm{D}}(\cdot)$	empirical L_τ-risk w.r.t. the data set D
$L_p(\mu)$	space of equivalence class of p-integrable functions w.r.t. μ

$W_p^\alpha(\mu)$	Sobolev space of order α
$B_{p,q}^\alpha(\mu)$	Besov space of smoothness α
H	reproducing kernel Hilbert space (RKHS)
H_γ	Gaussian RKHS
k	kernel
K	Gram matrix
k_γ	Gaussian RBF kernel
$\|\cdot\|_2$	Euclidean norm
$\|\cdot\|_\infty$	supremum norm
$\|\cdot\|_H$	norm of RKHS H
$\|\cdot\|_{H_\gamma}$	norm of Gaussian RKHS H_γ
$\|\cdot\|_{B_{p,q}^\alpha(\mu)}$	semi-norm of the Besov spaces $B_{p,q}^\alpha(\mu)$
$\|\cdot\|_{B_{p,q}^\alpha(\mu)}$	norm of the Besov spaces $B_{p,q}^\alpha(\mu)$
$\|\cdot\|_{W_{p^\alpha}(\mu)}$	norm of the Sobolev space $W_{p^\alpha}(\mu)$
$\langle\cdot,\cdot\rangle_H$	inner product in the Hilbert space H
α	smoothing parameter of the target function/first dual variable
β	second dual variable
λ	regularization parameter
γ	width of the Gaussian RBF kernel
δ	difference between new and old value of dual variable α
η	difference between new and old value of dual variable β
f_{D}	decision function produced by a leaning method
$f_{\mathrm{D},\lambda}$	empirical SVM decision function w.r.t. the data set D
$f_{\mathrm{D},\lambda,\gamma}$	empirical SVM decision function w.r.t. the data set D in Gaussian RKHS
$\widehat{f}$	clip decision function
$f_{\mathrm{P},\lambda}$	general SVM decision function w.r.t P
$f^*_{L_\tau,\mathrm{P}}, \mu^\star_\tau$	conditional τ-expectile function
$\mathcal{A}(\lambda)$	approximation error function for $\lambda > 0$
id	identity map