Mach Learn (2016) 103:103–136
DOI 10.1007/s10994-015-5540-x

Tikhonov, Ivanov and Morozov regularization for support vector machine learning

Luca Oneto[1] · **Sandro Ridella**[1] · **Davide Anguita**[2]

Received: 11 October 2013 / Accepted: 26 October 2015 / Published online: 22 December 2015

Abstract Learning according to the structural risk minimization principle can be naturally expressed as an Ivanov regularization problem. Vapnik himself pointed out this connection, when deriving an actual learning algorithm from this principle, like the well-known support vector machine, but quickly suggested to resort to a Tikhonov regularization schema, instead. This was, at that time, the best choice because the corresponding optimization problem is easier to solve and in any case, under certain hypothesis, the solutions obtained by the two approaches coincide. On the other hand, recent advances in learning theory clearly show that the Ivanov regularization scheme allows a more effective control of the learning hypothesis space and, therefore, of the generalization ability of the selected hypothesis. We prove in this paper the equivalence between the Ivanov and Tikhonov approaches and, for the sake of completeness, their connection to Morozov regularization, which has been show to be useful when effective estimation of the noise in the data is available. We also show that this equivalence is valid under milder conditions on the loss function with respect to Vapnik's original proposal. These results allows us to derive several methods for performing SRM learning according to an Ivanov or Morozov regularization scheme, but using Tikhonov-based solvers, which have been thoroughly studied in the last decades and for which very efficient implementations have been proposed.

Keywords Structural risk minimization · Tikhonov regularization · Ivanov regularization · Morozov regularization · Support vector machine

✉ Luca Oneto
Luca.Oneto@unige.it

Sandro Ridella
Sandro.Ridella@unige.it

Davide Anguita
Davide.Anguita@unige.it

[1] DITEN - University of Genoa, Via Opera Pia 11a, 16145 Genoa, Italy

[2] DIBRIS - University of Genoa, Via Opera Pia 13, 16145 Genoa, Italy

1 Introduction

The structural risk minimization (SRM) principle, formulated by Vapnik in the Statistical Learning Theory (SLT) framework (Vapnik 1998, 2000), requires a learning procedure to search for the hypothesis space that guarantees the best trade-off between its complexity and its fitting capabilities on the training samples (Vapnik 1998; Anguita et al. 2012). Consequently, according to the SRM principle, the control of the hypothesis space size assumes a central role in learning (Guyon et al. 2010; Anguita et al. 2011a). Vapnik's original approach to the derivation of an actual learning algorithm, like the support vector machine (SVM) (Cortes and Vapnik 1995; Vapnik 1998; Pontil and Verri 1998), consisted in implementing the SRM principle through an Ivanov regularization scheme (Vapnik 1998; Ivanov 1976). This is a logical approach because the Ivanov regularization framework allows to handle directly the two main forces guiding the SRM-based learning: on one hand, the minimization of the empirical risk and, on the other hand, the control of the hypothesis space, where the hypothesis minimizing the risk is chosen from. For the sake of brevity, we will refer to this formulation, when addressing the SVM learning algorithm, as the Ivanov-based SVM (I-SVM).

However, in his seminal works, Vapnik resorted to an alternative formulation, based on Tikhonov regularization, which quickly became very successful, due to its excellent performance in real-world problems, and which is commonly referred to as the SVM algorithm (to avoid any confusion we will make reference to this formulation as T-SVM) (Vapnik 1998; Tikhonov et al. 1977). The main argument in favor of this option is that the T-SVM learning problem is easier to solve than the I-SVM one: in fact, the amount of effective solvers for T-SVM, which appeared in the literature in the following decades, support this claim (Platt 1998, 1999; Keerthi et al. 2001; Shawe-Taylor and Sun 2011).

In this paper we will show that the Ivanov regularization approach is directly linked with one of the most powerful measure of the generalization ability of a learning algorithm: the Rademacher Complexity (Bartlett and Mendelson 2003; Koltchinskii 2006; Bartlett et al. 2005). In particular we will show that it is possible to bound these quantities even for the case of Tikhonov and Morozov regularization, but, in order the direct control of the generalization ability of the learning algorithm, we have to resort to the Ivanov formulation. Since the Tikhonov regularization scheme for SRM learning does not allow to directly control the size of the hypothesis space, this produces a soft mismatch in the theory (Shawe-Taylor et al. 1998; Bartlett 1998). This is usually considered an acceptable price to pay in order to foster the applicability of the SVM learning algorithm to practical problems. However, being able to carefully fine-tuning the complexity of the hypothesis space can lead, in the data-dependent SRM framework (Bartlett and Mendelson 2003; Koltchinskii 2006), to remarkable improvements in the quality of the identified SVM solution. This is especially true when dealing with difficult classification problems where, for example, only few high-dimensional samples are available to train a reliable and effective classifier (Anguita et al. 2011b, 2012). In this case, the requirement of the SRM principle of precisely considering a series of hypothesis spaces of increasing size, for identifying the optimal class of functions, becomes of paramount importance (Vapnik 1998; Duan et al. 2003). Therefore, a desirable objective would be to achieve the best of the two worlds: addressing the I-SVM learning problem by exploiting the efficiency of T-SVM solvers.

Furthermore, as showed by Pelckmans et al. (2004) for the particular case of Least Square SVM (LS-SVM) (Suykens and Vandewalle 1999), a third approach can be taken in account, based on Morozov regularization. The Morozov regularization schema, despite being seldom used in practice, has been shown to be effective when reliable estimations of the noise

afflicting the data are available, so it is worth considering as a further and alternative learning formulation (Morozov et al. 1984).

In reaching the above mentioned objectives, we propose in this work some more general results, which are valid for any convex loss function, including the SVM hinge loss as a particular case, and prove the equivalence between Tikhonov, Ivanov and Morozov regularization schemas in a general setting. Then, we apply our findings to the particular case of SVM classifiers and propose several ways to solve I-SVM and M-SVM learning problems through use of efficient T-SVM solvers.

The paper is organized as follows: after introducing the supervised learning framework in Sect. 2, we revise the formulations of Tikhonov, Ivanov and Morozov regularization approaches in Sect. 3. Then, in Sect. 4, we prove the equivalence of the regularization paths of the three approaches and derive some general properties relating the corresponding optimal solutions. In Sect. 6 we specialize our findings to the particular case of SVM training and, in Sect. 7, we show experimentally the advantages and disadvantages of using the well-known T-SVM Sequential Minimal Optimization (SMO) solver (Platt 1998; Keerthi et al. 2001; Keerthi and Gilbert 2002; Fan et al. 2005) for addressing both I-SVM and M-SVM problems. Finally, Sect. 8 summarizes some concluding remarks.

2 The supervised learning framework

We recall the standard framework of supervised learning, where the goal is to approximate a relationship between inputs from a set $\mathcal{X} \subseteq \mathbb{R}^d$, and outputs from a set $\mathcal{Y} \subseteq \mathbb{R}$. A special case of interest is the discrete case, where $\mathcal{Y} \equiv \{-1, +1\}$ (i.e. the binary classification problem). The relationship between inputs and outputs is encoded by a fixed, but unknown, probability distribution μ on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Each element $(\boldsymbol{x}, y) = \boldsymbol{z} \in \mathcal{Z}$ is defined as a labeled example: the training phase consists of a learning algorithm, which exploits a sequence $\mathcal{D}_n = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n\} \in \mathcal{Z}^n$ of labeled examples and returns a function $h : \mathcal{X} \to \mathbb{R}$ chosen from a fixed set $\mathcal{H}$ of possible hypotheses. The learning algorithm maps $(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$ to $\mathcal{H}$, and the accuracy in representing the hidden relationship μ is measured with reference to a loss function $\ell : \mathbb{R} \times \mathbb{R} \to [0, \infty)$.

For any $h \in \mathcal{H}$, we define the generalization error $L(h)$ as the expectation of $\ell(h(\boldsymbol{x}), y)$ with respect to μ, $L(h) = \mathbb{E}_\mu \ell(h(\boldsymbol{x}), y)$, where we assume that each labelled sample is generated according to μ. Our scope is to find the best $h \in \mathcal{H}$ for which $L(h)$ is minimum. Unfortunately, $L(h)$ cannot be computed since μ is unknown, but we can easily compute its empirical version $\hat{L}(h)$:

$$\hat{L}(h) = \frac{1}{n} \sum_{i=1}^{n} \ell\left(h(\boldsymbol{x}_i), y_i\right). \tag{1}$$

We focus in this paper on convex (Boyd and Vandenberghe 2004; Bauschke and Combettes 2011), and Lipschitz continuos (Goldstein 1977) loss functions only, as they are quite common and allow to solve complex learning tasks with effective and efficient approaches (Cortes and Vapnik 1995; Bartlett et al. 2006; Lee et al. 1998; Shawe-Taylor and Cristianini 2004; Suykens and Vandewalle 1999): examples are the well-known hinge (Cortes and Vapnik 1995) and logistic (Collins et al. 2002) loss functions. On the contrary, the use of a non-convex loss leads to NP-problems, which cannot be exactly solved for sample sets whose cardinality exceeds few tens of data (e.g. $n > 30$) (Anthony 2001; Feldman et al. 2009), but for which approximate solutions can be eventually found (Lawler and Wood 1966; Yuille and

Rangarajan 2003). As a matter of fact, if one has to cope with a non-convex loss, a convex relaxation is often used in order to reformulate the problem so to make it computationally tractable.

3 Tikhonov, Ivanov and Morozov regularization problems

We address a supervised learning framework, where the class of functions is parameterized as follows:

$$h(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}) + b, \tag{2}$$

$\boldsymbol{\phi} : \mathbb{R}^d \rightarrow \mathbb{R}^D$ is a mapping function, $\boldsymbol{w} \in \mathbb{R}^D$, and $b \in \mathbb{R}$. The naïve approach to learning, namely the Empirical Risk Minimization (ERM) (Vapnik 1998; Bousquet et al. 2004), consists in searching for the function h that minimizes the empirical error:

$$h : \quad \arg\min_{\boldsymbol{w},b} \hat{L}(h). \tag{3}$$

As Problem (3) is convex, local minima are avoided (Boyd and Vandenberghe 2004) even though the solution $\{\boldsymbol{w}, b\}$ is not unique, in general. Unfortunately ERM is well known to lead to a severe overfitting and then to poor performance in classifying new data, originated by the same distribution μ but previously unseen.

Alternatively, in order to avoid the overfitting issue that afflicts the ERM procedure, the Tikhonov regularization technique (Tikhonov et al. 1977) can be exploited, which was proposed to solve ill-posed problems (Bishop 1995):

$$h : \quad \arg\min_{\boldsymbol{w},b} \hat{L}(h) + \frac{\lambda}{2} \|\boldsymbol{w}\|^2 \quad \text{or} \quad \arg\min_{\boldsymbol{w},b} \frac{1}{2} \|\boldsymbol{w}\|^2 + C\hat{L}(h), \tag{4}$$

where $\|\boldsymbol{w}\|$ is the Euclidean norm of $\boldsymbol{w}$, and implements an underfitting tendency, so that the regularization parameter $\lambda \in [0, \infty)$, or equivalently $C = \frac{1}{\lambda} \in (0, \infty]$, balances the influence of the underfitting and the overfitting terms (Vorontsov 2010; Bousquet et al. 2004).

A consequence of this formulation is that λ implicitly defines the class of functions $\mathcal{H}$, from which the models $h(\boldsymbol{x})$ are selected by the optimization procedure (Tikhonov et al. 1977; Vapnik 1998), but the relation between the regularization parameter and the size of the hypothesis space is not evident at all.

Differently to the Tikhonov scheme, the method of quasi-solutions, originally proposed by Ivanov and also known as Ivanov regularization (Ivanov 1976), allows to explicitly control the size of $\mathcal{H}$ by upper bounding the square norm of the admissible hypotheses (Pelckmans et al. 2004; Vapnik 1998; Anguita et al. 2012):

$$\begin{aligned} h : \quad & \arg\min_{\boldsymbol{w},b} \quad \hat{L}(h) \\ & s.t. \quad \|\boldsymbol{w}\|^2 \leq w^2_{\text{MAX}}, \end{aligned} \tag{5}$$

by means of the regularization parameter $w^2_{\text{MAX}} \in [0, \infty)$.

It is worthwhile noting that the solution $\{\boldsymbol{w}^*, b^*\}$ for Problem (5) is not unique, in general (Boyd and Vandenberghe 2004). In order to eliminate such potential ambiguity, we can simply opt for the function $h(\boldsymbol{x})$ characterized by minimum $\|\boldsymbol{w}\|$, namely the simplest (smoothest) possible solution. In order to highlight this, without modifying the nature of the regularization procedure, we propose an equivalent formulation to Problem (5):

$$h: \quad \arg\min_{\boldsymbol{w},b} \quad \|\boldsymbol{w}\| \\ s.t. \quad h \in \mathcal{S} \\ \mathcal{S} = \left\{ h : \hat{L}(h) = \arg\min_{\boldsymbol{w},b} \hat{L}(h) \; s.t. \; \|\boldsymbol{w}\|^2 \leq w_{\text{MAX}}^2 \right\}. \tag{6}$$

In order to simplify the notation of Problem (6) we simply add $\|\boldsymbol{w}\|$ to the argument of the minimum in Problem (5):

$$h: \quad \arg\min_{\boldsymbol{w},b,\|\boldsymbol{w}\|} \quad \hat{L}(h) \\ s.t. \quad \|\boldsymbol{w}\|^2 \leq w_{\text{MAX}}^2. \tag{7}$$

A third way to write our regularization problem is the less-known approach proposed by Morozov (Morozov et al. 1984; Pelckmans et al. 2004):

$$h: \quad \arg\min_{\boldsymbol{w},b} \quad \frac{1}{2}\|\boldsymbol{w}\|^2 \\ s.t. \quad \hat{L}(h) \leq \hat{L}_{\text{MAX}}. \tag{8}$$

In this case, the size of the hypothesis space is implicitly controlled by imposing an upper bound to the empirical error, namely $\hat{L}_{\text{MAX}} \in [0, \infty)$.

It is worthwhile noting that also the solution $\{\boldsymbol{w}^*, b^*\}$ for Problem (8) is not unique, in general (Boyd and Vandenberghe 2004). In order to eliminate such potential ambiguity, we can simply opt for the function $h(\boldsymbol{x})$ characterized by minimum $\hat{L}(h)$, namely the solution with minimum error. In order to highlight this, without modifying the nature of the regularization procedure, we propose an equivalent formulation to Problem (8):

$$h: \quad \arg\min_{\boldsymbol{w},b} \quad \hat{L}(h) \\ s.t. \quad h \in \mathcal{S} \\ \mathcal{S} = \left\{ h : \|\boldsymbol{w}\|^2 = \arg\min_{\boldsymbol{w},b} \|\boldsymbol{w}\|^2 \; s.t. \; \hat{L}(h) \leq \hat{L}_{\text{MAX}}. \right\}. \tag{9}$$

In order to simplify the notation of Problem (9) we simply add $\hat{L}(h)$ to the argument of the minimum in Problem (8):

$$h: \quad \arg\min_{\boldsymbol{w},b,\hat{L}(h)} \quad \frac{1}{2}\|\boldsymbol{w}\|^2 \\ s.t. \quad \hat{L}(h) \leq \hat{L}_{\text{MAX}}. \tag{10}$$

The philosophy underlying the Morozov regularization approach consists in choosing the simplest function, by minimizing $\|\boldsymbol{w}\|^2$, which performs better than a pre-determined performance threshold on the training set. Clearly, if the threshold $\hat{L}_{\text{MAX}}$ is too small, a solution could not exist: therefore, for the sake of simplicity, we will assume in the rest of the paper that $\hat{L}_{\text{MAX}}$ is large enough so that a solution can be found. This hypothesis does not modify the nature of Morozov regularization, while it helps simplifying the subsequent analysis.

It is important to note that the Representer Theorem holds for all the previous regularization approaches (Aronszajn 1951; Schölkopf et al. 2001; Dinuzzo and Schölkopf 2012). Consequently, the solution to the Tikhonov, Ivanov and Morozov optimization problems can be expressed as:

$$\boldsymbol{w} = \sum_{i=1}^{n} p_i \boldsymbol{\phi}(\boldsymbol{x}_i), \tag{11}$$

where $p_i \in \mathbb{R}\ \forall i \in \{1, \ldots, n\}$. Because of this property, the solution function $h(\boldsymbol{x})$ can be written as:

$$h(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}) + b = \sum_{i=1}^{n} p_i \boldsymbol{\phi}(\boldsymbol{x}_i) \cdot \boldsymbol{\phi}(\boldsymbol{x}) + b = \sum_{i=1}^{n} p_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b \tag{12}$$

where we made use of the well-known kernel trick (Berlinet and Thomas-Agnan 2004; Vapnik 1998; Schölkopf 2001; Shawe-Taylor and Cristianini 2004) and $K(\cdot, \cdot)$ is the kernel function.

4 A general approach for solving Ivanov and Morozov problems through the Tikhonov formulation

Although, according to the SRM framework, learning can be easily implemented by an Ivanov regularization approach, a Tikhonov formulation has been usually preferred as it is easier to solve, and several effective methods have been developed throughout the years for this purpose (Platt 1998, 1999; Keerthi et al. 2001; Shawe-Taylor and Sun 2011). In this section, we show that the Tikhonov, Ivanov and Morozov regularization approaches are three faces of the same problem: in particular, we will show how the Ivanov and Morozov problems can be solved through the procedures originally designed for the Tikhonov based formulation.

4.1 Equivalence of Tikhonov, Ivanov and Morozov formulations

At first, we show that a value of the Tikhonov regularization parameter exists such that the three problems are equivalent.[1]

Theorem 1 *Let us consider an Ivanov (or Morozov) regularization problem, as formulated in Eqs.* (7) *and* (10)*; then, there exists a value of* $C = \frac{1}{\lambda}$ *for the Tikhonov regularization Problem* (4) *such that the formulations are equivalent.*

Proof As a first step, let us consider the Ivanov Problem (7). Because of its convexity, we can compute the Lagrange dual function and solve the associate optimization problem (Boyd and Vandenberghe 2004):

$$\begin{aligned} \left(\boldsymbol{w}_I^*, b_I^*, \left\|\boldsymbol{w}_I^*\right\|, \lambda_I^*\right): \quad & \arg \min_{\boldsymbol{w}, b, \|\boldsymbol{w}\|} \max_{\lambda \geq 0} \ \hat{L}(h) + \frac{\lambda}{2}\left(\|\boldsymbol{w}\|^2 - w_{\mathrm{MAX}}^2\right) \\ & s.t. \ \ \lambda\left(\|\boldsymbol{w}\|^2 - w_{\mathrm{MAX}}^2\right) = 0, \end{aligned} \tag{13}$$

where λ is the Lagrange multiplier of the constraint on the class of functions. Then, if we exploit in the Tikhonov problem of Eq. (4) the value of λ_I^*, obtained by the minimization of the dual function of the Ivanov regularization problem shown above, we obtain:

$$\begin{aligned} \left(\boldsymbol{w}_T^*, b_T^*\right): \quad & \arg \min_{\boldsymbol{w}, b} \ \hat{L}(h) + \frac{\lambda_I^*}{2}\|\boldsymbol{w}\|^2 = \\ & \arg \min_{\boldsymbol{w}, b} \ \hat{L}(h) + \frac{\lambda_I^*}{2}\left(\|\boldsymbol{w}\|^2 - w_{\mathrm{MAX}}^2\right), \end{aligned} \tag{14}$$

[1] From here further, we set $\hat{L}(h) = \hat{L}$ for the sake of notational simplicity.

since w^2_{MAX} is constant with respect to the minimization problem. As Problems (13) and (14) are equal, they also have the same solution and then a value of $\lambda = \frac{1}{C}$ exists such that the two formulations are equivalent.

Concerning the Morozov approach, the proof is analogous. We can compute the Lagrange dual function of Problem (10):

$$\begin{aligned}\left(\boldsymbol{w}^*_M, b^*_M, C^*_M, \hat{L}^*_M\right): \quad \arg \min_{\boldsymbol{w},b,\hat{L}(h)} \max_{C\geq 0} \quad & \frac{1}{2}\|\boldsymbol{w}\|^2 + C\left(\hat{L}(h) - \hat{L}_{\text{MAX}}\right) \\ s.t. \quad & \hat{L}(h) \leq \hat{L}_{\text{MAX}} \\ & C\left(\hat{L}(h) - \hat{L}_{\text{MAX}}\right) = 0. \end{aligned} \tag{15}$$

Then, if we exploit the optimal value of the Lagrange multiplier C^*_M, obtained through the previous problem, in the Tikhonov formulation of Eq. (4), we obtain:

$$\begin{aligned}\left(\boldsymbol{w}^*_T, b^*_T\right): \quad \arg\min_{\boldsymbol{w},b} \quad & \frac{1}{2}\|\boldsymbol{w}\|^2 + C^*_M \hat{L}(h) = \\ \arg\min_{\boldsymbol{w},b} \quad & \frac{1}{2}\|\boldsymbol{w}\|^2 + C^*_M\left(\hat{L}(h) - \hat{L}_{\text{MAX}}\right). \end{aligned} \tag{16}$$

Problems (15) and (16) are consequently equal, thus they have the same solution, i.e. there exists a value of $C = \frac{1}{\lambda}$ such that the two formulations are equivalent. □

Note that it is possible to find the same results, in a more general framework, in Bauschke and Combettes (2011). The following theorems allow to prove that the solutions of the Ivanov–Tikhonov and the Morozov–Tikhonov approaches coincide.

Theorem 2 *Let us consider the Tikhonov and Ivanov formulations. Let* $\left(\left\|\boldsymbol{w}^*_T\right\|, \hat{L}^*_T\right)$ *and* $\left(\left\|\boldsymbol{w}^*_I\right\|, \hat{L}^*_I\right)$ *be the solutions of, respectively, the Tikhonov and the Ivanov problem. If* $\left\|\boldsymbol{w}^*_T\right\| = \left\|\boldsymbol{w}^*_I\right\|$ *for a given* $C = \frac{1}{\lambda}$ *and for a given* w_{MAX}*, then* $\hat{L}^*_T = \hat{L}^*_I$ *and vice-versa.*

Proof Based on the definition of minimum of the Tikhonov problem, from Eq. (4) we have that

$$\frac{1}{2}\left(\left\|\boldsymbol{w}^*_I\right\|\right)^2 + C\hat{L}^*_I \geq \frac{1}{2}\left(\left\|\boldsymbol{w}^*_T\right\|\right)^2 + C\hat{L}^*_T. \tag{17}$$

As we supposed that C is such that $\left\|\boldsymbol{w}^*_T\right\| = \left\|\boldsymbol{w}^*_I\right\|$, then:

$$\hat{L}^*_I \geq \hat{L}^*_T. \tag{18}$$

However, as $\hat{L}^*_I$ is the solution to Problem (7), we must have that

$$\hat{L}^*_I = \hat{L}^*_T. \tag{19}$$

If, instead, we suppose that C is such that $\hat{L}^*_I = \hat{L}^*_T$, then:

$$\left\|\boldsymbol{w}^*_I\right\| \geq \left\|\boldsymbol{w}^*_T\right\|. \tag{20}$$

However, since in Problem (7) we supposed to search for the solution characterized by minimum $\|\boldsymbol{w}\|$, we have:

$$\left\|\boldsymbol{w}^*_I\right\| = \left\|\boldsymbol{w}^*_T\right\|. \tag{21}$$

□

Note that, if we did not opt for the smoothest solution, it would be impossible to prove this last property. In fact, a counterexample where different values of $\|\boldsymbol{w}\|$ allow to achieve the same minimum $\hat{L}_T^*$ can be found in Anguita et al. (2011b).

Theorem 3 *Let us consider the Tikhonov and Morozov formulations. Let* $\left(\left\|\boldsymbol{w}_T^*\right\|, \hat{L}_T^*\right)$ *and* $\left(\left\|\boldsymbol{w}_M^*\right\|, \hat{L}_M^*\right)$ *be the solutions of, respectively, the Tikhonov and the Morozov problems. If* $\left\|\boldsymbol{w}_T^*\right\| = \left\|\boldsymbol{w}_M^*\right\|$ *for a given* $C = \frac{1}{\lambda}$ *and for a given* $\hat{L}_{MAX}$*, then* $\hat{L}_T^* = \hat{L}_M^*$ *and vice-versa.*

Proof Based on the definition of minimum of the Tikhonov problem, from Eq. (4) we have that

$$\frac{1}{2}\left(\left\|\boldsymbol{w}_M^*\right\|\right)^2 + C\hat{L}_M^* \geq \frac{1}{2}\left(\left\|\boldsymbol{w}_T^*\right\|\right)^2 + C\hat{L}_T^*. \tag{22}$$

As we supposed that λ is such that $\hat{L}_T^* = \hat{L}_M^*$, then:

$$\left\|\boldsymbol{w}_M^*\right\| \geq \left\|\boldsymbol{w}_T^*\right\|, \tag{23}$$

However, as $\hat{L}_M^*$ is the solution to Problem (10), we must have that

$$\left\|\boldsymbol{w}_M^*\right\| = \left\|\boldsymbol{w}_T^*\right\|. \tag{24}$$

If instead we suppose that λ is such that $\left\|\boldsymbol{w}_M^*\right\| = \left\|\boldsymbol{w}_T^*\right\|$, then:

$$\hat{L}_M^* \geq \hat{L}_T^*. \tag{25}$$

In particular, since in problem Problem (10) we force $\left\|\boldsymbol{w}_M^*\right\| = \left\|\boldsymbol{w}_T^*\right\|$, we have that $\hat{L}_M^*$ is forced to be as small as possible. Consequently $\hat{L}_M^* > \hat{L}_T^*$ is not possible and we have that:

$$\hat{L}_M^* = \hat{L}_T^*. \tag{26}$$

□

4.2 Solving Ivanov and Morozov problems with Tikhonov solvers

In the following, we prove some properties that allow us to define general procedures for solving an Ivanov or a Morozov problem through the techniques designed for Tikhonov formulations. We start by depicting the behavior of the Tikhonov problem solution as the regularization parameter C grows. The following theorem is propaedeutic to the derivation of such results.

Theorem 4 *Let us consider the Tikhonov formulation. Let us solve Problem* (4) *for two given values of the regularization parameter* C_1 *and* $C_2 > C_1$*. In particular, let the solutions of the problem be, respectively,* $\left(\left\|\boldsymbol{w}_{C_1}^*\right\|, \hat{L}_{C_1}^*\right)$ *for* C_1 *and* $\left(\left\|\boldsymbol{w}_{C_2}^*\right\|, \hat{L}_{C_2}^*\right)$ *for* C_2 *so that the corresponding values of the objective functions are:*

$$K^{C_1} = \frac{1}{2}\left(\left\|\boldsymbol{w}_{C_1}^*\right\|\right)^2 + C_1\hat{L}_{C_1}^*, \quad K^{C_2} = \frac{1}{2}\left(\left\|\boldsymbol{w}_{C_2}^*\right\|\right)^2 + C_2\hat{L}_{C_2}^*. \tag{27}$$

Then:

$$K^{C_2} \geq K^{C_1}. \tag{28}$$

Proof Since $C_2 > C_1$ $\forall (\boldsymbol{w}, b) \in \mathbb{R}^D \times \mathbb{R}$ we have that:

$$\frac{1}{2}\|\boldsymbol{w}\|^2 + C_1 \hat{L}(h) \leq \frac{1}{2}\|\boldsymbol{w}\|^2 + C_2 \hat{L}(h). \tag{29}$$

The statement follows from taking the infimum over $(\boldsymbol{w}, b)$ on both sides. □

By exploiting the result of the previous theorem, we prove the following property, which we will exploit in the following to design actual learning algorithms.

Theorem 5 *Let us consider the Tikhonov formulation. Given $C_1, C_2 \in [0, +\infty]$ such that $C_2 > C_1$, let us solve Problem (4) and let K^{C_1} and K^{C_2} be the corresponding values of the objective functions, then:*

$$\left(\left\|\boldsymbol{w}^*_{C_2}\right\| > \left\|\boldsymbol{w}^*_{C_1}\right\| \implies \hat{L}^*_{C_2} < \hat{L}^*_{C_1}\right) \vee \left(\left\|\boldsymbol{w}^*_{C_2}\right\| = \left\|\boldsymbol{w}^*_{C_1}\right\| \implies \hat{L}^*_{C_2} = \hat{L}^*_{C_1}\right). \tag{30}$$

Proof In this proof, we proceed by considering all the possible cases, proving by contradiction that configurations other than the ones of the thesis are not admissible.

As a first step, suppose $\left\|\boldsymbol{w}^*_{C_2}\right\| < \left\|\boldsymbol{w}^*_{C_1}\right\|$. If $\hat{L}^*_{C_2} < \hat{L}^*_{C_1}$, then $K^{C_2} < K^{C_1}$, which is impossible (see Theorem 4). If $\hat{L}^*_{C_2} = \hat{L}^*_{C_1}$, then:

$$K'^{C_1} = \frac{1}{2}\left(\left\|\boldsymbol{w}^*_{C_2}\right\|\right)^2 + C_1 \hat{L}^*_{C_2} < K^{C_1}, \tag{31}$$

which contradicts the hypothesis that K^{C_1} is the global minimum and, then, is not admissible.

If $\hat{L}^*_{C_2} > \hat{L}^*_{C_1}$, then:

$$K'^{C_1} = \frac{1}{2}\left(\left\|\boldsymbol{w}^*_{C_2}\right\|\right)^2 + C_1 \hat{L}^*_{C_2} \geq \frac{1}{2}\left(\left\|\boldsymbol{w}^*_{C_1}\right\|\right)^2 + C_1 \hat{L}^*_{C_1} = K^{C_1}. \tag{32}$$

From Eq. (32), we get:

$$C_1 \geq \frac{\left(\left\|\boldsymbol{w}^*_{C_1}\right\|\right)^2 - \left(\left\|\boldsymbol{w}^*_{C_2}\right\|\right)^2}{2\left(\hat{L}^*_{C_2} - \hat{L}^*_{C_1}\right)}. \tag{33}$$

Analogously, we have for K^{C_2}:

$$\frac{1}{2}\left(\left\|\boldsymbol{w}^*_{C_1}\right\|\right)^2 + C_2 \hat{L}^*_{C_1} \geq \frac{1}{2}\left(\left\|\boldsymbol{w}^*_{C_2}\right\|\right)^2 + C_2 \hat{L}^*_{C_2} = K^{C_2}, \tag{34}$$

from which we obtain:

$$C_2 \leq \frac{\left(\left\|\boldsymbol{w}^*_{C_1}\right\|\right)^2 - \left(\left\|\boldsymbol{w}^*_{C_2}\right\|\right)^2}{2\left(\hat{L}^*_{C_2} - \hat{L}^*_{C_1}\right)} \tag{35}$$

By joining Eqs. (33) and (35), we have that $C_2 < C_1$, which contradicts the hypotheses.

Suppose now that $\left\|\boldsymbol{w}^*_{C_2}\right\| = \left\|\boldsymbol{w}^*_{C_1}\right\|$. If $\hat{L}^*_{C_2} < \hat{L}^*_{C_1}$, then

$$K'^{C_1} = \frac{1}{2}\left(\left\|\boldsymbol{w}^*_{C_1}\right\|\right)^2 + C_1 \hat{L}^*_{C_2} < K^{C_1} \tag{36}$$

which is impossible, as we supposed that K^{C_1} is the global minimum. Analogously, if $\hat{L}^*_{C_2} > \hat{L}^*_{C_1}$, then:

$$K'^{C_2} = \frac{1}{2}\left(\left\|\boldsymbol{w}^*_{C_2}\right\|\right)^2 + C_2 \hat{L}^*_{C_1} < K^{C_2}, \tag{37}$$

or, in other words, K^{C_2} is not the global minimum. This contradicts the hypothesis and, thus, this configuration is not admissible.

Finally, let us consider $\left\|\boldsymbol{w}^*_{C_2}\right\| > \left\|\boldsymbol{w}^*_{C_1}\right\|$. If $\hat{L}^*_{C_2} = \hat{L}^*_{C_1}$, then

$$K'^{C_2} = \frac{1}{2}\left(\left\|\boldsymbol{w}^*_{C_1}\right\|\right)^2 + C_2\hat{L}^*_{C_2} < K^{C_2} \tag{38}$$

which is, again, impossible as K^{C_2} is supposed to be the global minimum. As a last step, let us consider the case $\hat{L}^*_{C_2} > \hat{L}^*_{C_1}$, for which we have:

$$K'^{C_2} = \frac{1}{2}\left(\left\|\boldsymbol{w}^*_{C_1}\right\|\right)^2 + C_2\hat{L}^*_{C_1} < K^{C_2}, \tag{39}$$

which violates the hypotheses of the theorem. Thus, all the configurations of $\left\|\boldsymbol{w}^*_{C_{1,2}}\right\|$ and $\hat{L}^*_{C_{1,2}}$ other than the ones of the thesis are not admissible. □

The following theorem proves that, if $\left\|\boldsymbol{w}^*_C\right\|$ stops increasing as C increases, it will remain the same, regardless of the value assumed by the regularization parameter.

Theorem 6 *Let us consider the Tikhonov formulation. Let* $\left\|\boldsymbol{w}^*_{C_\infty}\right\|$ *be the solution to the regularization problem for a given value of* C_∞. *If* $\exists C > C_\infty$ *such that* $\left\|\boldsymbol{w}^*_C\right\| = \left\|\boldsymbol{w}^*_{C_\infty}\right\|$, *then* $\left\|\boldsymbol{w}^*_C\right\|$ *will not vary* $\forall C \geq C_\infty$.

Proof Let us consider a value C_1 of the regularization parameter such that $C_1 < C_\infty < C < \infty$ and $\left\|\boldsymbol{w}^*_C\right\| = \left\|\boldsymbol{w}^*_{C_\infty}\right\| > \left\|\boldsymbol{w}^*_{C_1}\right\|$. Then, we have that $\hat{L}^*_C = \hat{L}(h^*_{C_\infty}) < \hat{L}(h^*_{C_1})$. Moreover let us consider $C_2 \in [C_1, C_\infty]$. Thanks to the convexity of the problem, we have that $\forall\alpha \in [0, 1]$:

$$\hat{L}(h_\alpha) = \hat{L}\left(h^*_{C_1}(1-\alpha) + h^*_{C_\infty}\alpha\right) \leq (1-\alpha)\hat{L}\left(h^*_{C_1}\right) + \alpha\hat{L}\left(h^*_{C_\infty}\right). \tag{40}$$

Then we can define another Tikhonov problem, whose solution is constrained on the line that connects $h^*_{C_1}$ and $h^*_{C_\infty}$ and which can be parametrized as $h_\alpha = h^*_{C_1}(1-\alpha) + h^*_{C_\infty}\alpha$ ($\boldsymbol{w}_\alpha = (1-\alpha)\boldsymbol{w}^*_{C_1} + \alpha\boldsymbol{w}^*_{C_\infty}$) with $\alpha \in [0, 1]$. Then restriction of Problem (4) on the segment joining $\boldsymbol{w}^*_{C_1}$ and $\boldsymbol{w}^*_{C_\infty}$ can be formulated as:

$$\min_\alpha \quad \frac{1}{2}\|\boldsymbol{w}_\alpha\|^2 + C\hat{L}(h_\alpha). \tag{41}$$

Based on these considerations it is easy to show that (see also Fig. 1):

$$\begin{aligned}
&\min_{\boldsymbol{w},b} \quad \frac{1}{2}\|\boldsymbol{w}\|^2 + C_2\hat{L}(h) && (42)\\
&\leq \min_\alpha \quad \frac{1}{2}\|\boldsymbol{w}_\alpha\|^2 + C_2\hat{L}(h_\alpha) && (43)\\
&\leq \min_\alpha \quad \frac{1}{2}\left\|(1-\alpha)\boldsymbol{w}^*_{C_1} + \alpha\boldsymbol{w}^*_{C_\infty}\right\|^2 \\
&\quad + C_2\left[(1-\alpha)\hat{L}\left(h^*_{C_1}\right) + \alpha\hat{L}\left(h^*_{C_\infty}\right)\right] && (44)\\
&= \min_\alpha \quad \frac{1}{2}\left\|\boldsymbol{w}^*_{C_1} + \alpha\left(\boldsymbol{w}^*_{C_\infty} - \boldsymbol{w}^*_{C_1}\right)\right\|^2 \\
&\quad + C_2\left[\hat{L}\left(h^*_{C_1}\right) + \alpha\left(\hat{L}\left(h^*_{C_\infty}\right) - \hat{L}\left(h^*_{C_1}\right)\right)\right] && (45)
\end{aligned}$$

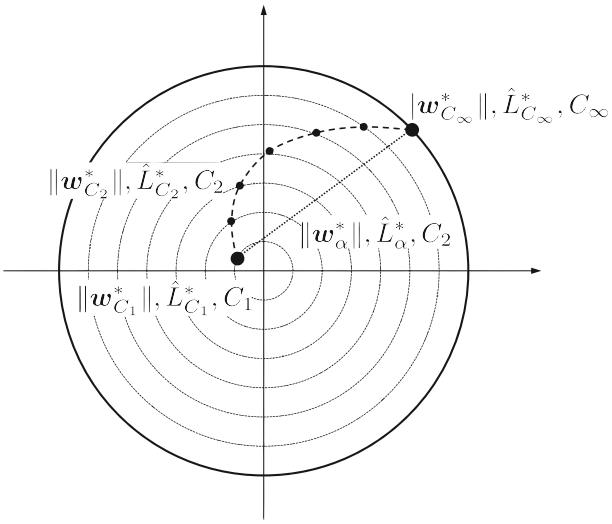

Fig. 1 Reformulation of the Tikhonov problem of Eq. (4) for proving Theorem 6

$$= \min_{\alpha} \quad \frac{1}{2}\left[\left\|\boldsymbol{w}^*_{C_1}\right\|^2 + \alpha^2 \left\|\boldsymbol{w}^*_{C_\infty} - \boldsymbol{w}^*_{C_1}\right\|^2 + 2\alpha \boldsymbol{w}^*_{C_1} \cdot \left(\boldsymbol{w}^*_{C_\infty} - \boldsymbol{w}^*_{C_1}\right)\right] + C_2\left[\hat{L}\left(h^*_{C_1}\right) + \alpha\left(\hat{L}\left(h^*_{C_\infty}\right) - \hat{L}\left(h^*_{C_1}\right)\right)\right] \tag{46}$$

This last minimization problem (Eq. (46)) can be easily solved:

$$\frac{d}{d\alpha}\left\{\frac{1}{2}\left[\left\|\boldsymbol{w}^*_{C_1}\right\|^2 + \alpha^2 \left\|\boldsymbol{w}^*_{C_\infty} - \boldsymbol{w}^*_{C_1}\right\|^2 + 2\alpha \boldsymbol{w}^*_{C_1} \cdot \left(\boldsymbol{w}^*_{C_\infty} - \boldsymbol{w}^*_{C_1}\right)\right] + C_2\left[\hat{L}\left(h^*_{C_1}\right) + \alpha\left(\hat{L}\left(h^*_{C_\infty}\right) - \hat{L}\left(h^*_{C_1}\right)\right)\right]\right\} \tag{47}$$

$$= \alpha \left\|\boldsymbol{w}^*_{C_\infty} - \boldsymbol{w}^*_{C_1}\right\|^2 + \boldsymbol{w}^*_{C_1} \cdot \left(\boldsymbol{w}^*_{C_\infty} - \boldsymbol{w}^*_{C_1}\right) + C_2\left[\hat{L}\left(h^*_{C_\infty}\right) - \hat{L}\left(h^*_{C_1}\right)\right] = 0 \tag{48}$$

$$\alpha = \frac{\boldsymbol{w}^*_{C_1} \cdot \left(\boldsymbol{w}^*_{C_1} - \boldsymbol{w}^*_{C_\infty}\right) + C_2\left[\hat{L}\left(h^*_{C_1}\right) - \hat{L}\left(h^*_{C_\infty}\right)\right]}{\left\|\boldsymbol{w}^*_{C_\infty} - \boldsymbol{w}^*_{C_1}\right\|^2} \tag{49}$$

Then we obtain that:

$$\min_{\alpha} \quad \frac{1}{2}\left[\left\|\boldsymbol{w}^*_{C_1}\right\|^2 + \alpha^2 \left\|\boldsymbol{w}^*_{C_\infty} - \boldsymbol{w}^*_{C_1}\right\|^2 + 2\alpha \boldsymbol{w}^*_{C_1} \cdot \left(\boldsymbol{w}^*_{C_\infty} - \boldsymbol{w}^*_{C_1}\right)\right] + C_2\left[\hat{L}\left(h^*_{C_1}\right) + \alpha\left(\hat{L}\left(h^*_{C_\infty}\right) - \hat{L}\left(h^*_{C_1}\right)\right)\right] \tag{50}$$

$$= \frac{1}{2}\left\|\boldsymbol{w}^*_{C_1}\right\|^2 + C_2\hat{L}\left(h^*_{C_1}\right) - \frac{1}{2}\alpha^2 \left\|\boldsymbol{w}^*_{C_\infty} - \boldsymbol{w}^*_{C_1}\right\|^2 \tag{51}$$

Based on these last results, we derive that:

$$\min_{\boldsymbol{w},b} \quad \frac{1}{2}\|\boldsymbol{w}\|^2 + C_2\hat{L}(h) \tag{52}$$

$$\leq \frac{1}{2}\left\|\boldsymbol{w}^*_{C_1}\right\|^2 + C_2\hat{L}\left(h^*_{C_1}\right) - \frac{1}{2}\alpha^2 \left\|\boldsymbol{w}^*_{C_\infty} - \boldsymbol{w}^*_{C_1}\right\|^2 \tag{53}$$

since $\frac{1}{2}\alpha^2 \left\| \boldsymbol{w}^*_{C_\infty} - \boldsymbol{w}^*_{C_1} \right\|^2 \geq 0$. Now we can observe that:

$$C_2 = C_1 \rightarrow \alpha = 0 \rightarrow \frac{\boldsymbol{w}^*_{C_1} \cdot \left(\boldsymbol{w}^*_{C_1} - \boldsymbol{w}^*_{C_\infty} \right) + C_2 \left[\hat{L}\left(h^*_{C_1} \right) - \hat{L}\left(h^*_{C_\infty} \right) \right]}{\left\| \boldsymbol{w}^*_{C_\infty} - \boldsymbol{w}^*_{C_1} \right\|^2} = 0$$

$$\boldsymbol{w}^*_{C_1} \cdot \left(\boldsymbol{w}^*_{C_1} - \boldsymbol{w}^*_{C_\infty} \right) = -C_1 \left[\hat{L}\left(h^*_{C_1} \right) - \hat{L}\left(h^*_{C_\infty} \right) \right] \tag{54}$$

$$C_2 = C_\infty \rightarrow \alpha = 1 \rightarrow \frac{\boldsymbol{w}^*_{C_1} \cdot \left(\boldsymbol{w}^*_{C_1} - \boldsymbol{w}^*_{C_\infty} \right) + C_2 \left[\hat{L}\left(h^*_{C_1} \right) - \hat{L}\left(h^*_{C_\infty} \right) \right]}{\left\| \boldsymbol{w}^*_{C_\infty} - \boldsymbol{w}^*_{C_1} \right\|^2} = 1$$

$$\boldsymbol{w}^*_{C_1} \cdot \left(\boldsymbol{w}^*_{C_1} - \boldsymbol{w}^*_{C_\infty} \right) = -C_\infty \left[\hat{L}\left(h^*_{C_1} \right) - \hat{L}\left(h^*_{C_\infty} \right) \right] + \left\| \boldsymbol{w}^*_{C_\infty} - \boldsymbol{w}^*_{C_1} \right\|^2 \tag{55}$$

Based on these two limit cases we can observe that $\forall C_2 \in (C_1, C_\infty)$ α falls, as hypothesized, in $(0, 1)$. Let us take α

$$\alpha = \frac{\boldsymbol{w}^*_{C_1} \cdot \left(\boldsymbol{w}^*_{C_1} - \boldsymbol{w}^*_{C_\infty} \right) + C_2 \left[\hat{L}\left(h^*_{C_1} \right) - \hat{L}\left(h^*_{C_\infty} \right) \right]}{\left\| \boldsymbol{w}^*_{C_\infty} - \boldsymbol{w}^*_{C_1} \right\|^2}, \tag{56}$$

by exploiting Eq. (54) we have that

$$\alpha = \frac{-C_1 \left[\hat{L}\left(h^*_{C_1} \right) - \hat{L}\left(h^*_{C_\infty} \right) \right] + C_2 \left[\hat{L}\left(h^*_{C_1} \right) - \hat{L}\left(h^*_{C_\infty} \right) \right]}{\left\| \boldsymbol{w}^*_{C_\infty} - \boldsymbol{w}^*_{C_1} \right\|^2} > 0. \tag{57}$$

By exploiting, instead, Eq. (55) we have that

$$\frac{-C_\infty \left[\hat{L}\left(h^*_{C_1} \right) - \hat{L}\left(h^*_{C_\infty} \right) \right] + \left\| \boldsymbol{w}^*_{C_\infty} - \boldsymbol{w}^*_{C_1} \right\|^2 + C_2 \left[\hat{L}\left(h^*_{C_1} \right) - \hat{L}\left(h^*_{C_\infty} \right) \right]}{\left\| \boldsymbol{w}^*_{C_\infty} - \boldsymbol{w}^*_{C_1} \right\|^2} =$$

$$1 - \frac{(C_\infty - C_1) \left[\hat{L}\left(h^*_{C_1} \right) - \hat{L}\left(h^*_{C_\infty} \right) \right]}{\left\| \boldsymbol{w}^*_{C_\infty} - \boldsymbol{w}^*_{C_1} \right\|^2} < 1 \tag{58}$$

From this last properties it is possible to state that $\forall C_2 \in (C_1, C_\infty)$ we have that:

$$\left\| \boldsymbol{w}^*_{C_1} \right\| < \left\| \boldsymbol{w}^*_{C_2} \right\| < \left\| \boldsymbol{w}^*_{C_\infty} \right\|. \tag{59}$$

Consequently

$$\hat{L}(h^*_{C_1}) > \hat{L}(h^*_{C_2}) > \hat{L}(h^*_{C_\infty}). \tag{60}$$

This means that if $\exists C_3 > C > C_\infty$ such that $\left\| \boldsymbol{w}^*_{C_3} \right\| > \left\| \boldsymbol{w}^*_{C_\infty} \right\|$ it is not possible that $\left\| \boldsymbol{w}^*_{C} \right\| = \left\| \boldsymbol{w}^*_{C_\infty} \right\|$ because in this case $\forall C \in (C_\infty, C_3)$ we must have that:

$$\left\| \boldsymbol{w}^*_{C_\infty} \right\| < \left\| \boldsymbol{w}^*_{C} \right\| < \left\| \boldsymbol{w}^*_{C_3} \right\|. \tag{61}$$

Consequently the statement of the Theorem is proved.

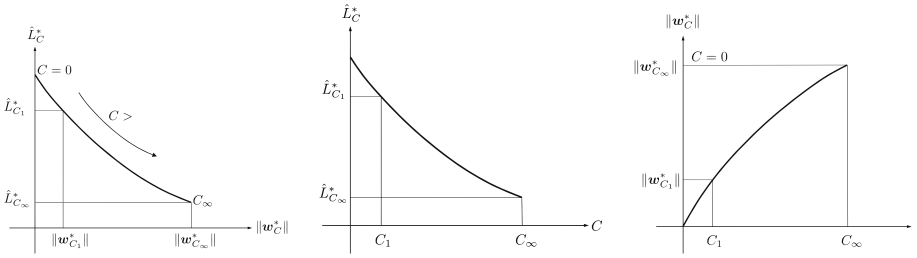

Fig. 2 Results of the Theorem 6: the monotonicity of the solution path

A particular case is when there is no C_1, which satisfies the property discussed in the proof ($C_1 < C_\infty < C < \infty$ and $\left\|\boldsymbol{w}^*_C\right\| = \left\|\boldsymbol{w}^*_{C_\infty}\right\| > \left\|\boldsymbol{w}^*_{C_1}\right\|$). In this case we have to suppose, by contradiction, that $\exists C_2$ such that $C_\infty < C < C_2 < \infty$ and $\left\|\boldsymbol{w}^*_C\right\| = \left\|\boldsymbol{w}^*_{C_\infty}\right\| < \left\|\boldsymbol{w}^*_{C_2}\right\|$. Then, by using the same argument exploited above, it is possible to show that $\left\|\boldsymbol{w}^*_C\right\| < \left\|\boldsymbol{w}^*_{C_\infty}\right\| < \left\|\boldsymbol{w}^*_{C_2}\right\|$, which contradict the hypothesis. Therefore $\left\|\boldsymbol{w}^*_C\right\| = \left\|\boldsymbol{w}^*_{C_\infty}\right\| = \left\|\boldsymbol{w}^*_{C_2}\right\|$. We did not report the details because they are analogous to the one reported before. □

A very familiar case where the assumptions of Theorem 6 are met is the one of the SVM with the hinge loss function: in Hastie et al. (2004) the entire regularization path is studied. When $C = 0$, the solution is $\boldsymbol{w} = \boldsymbol{0}$, then there is a C_∞ for which the solution will not change $\forall C \geq C_\infty$.

From the previous results, it can be clearly noted that, if the minimum value of the empirical error has not been reached yet, the empirical error $\hat{L}\left(h^*_C\right)$ monotonically decreases (and the regularization term $\left\|\boldsymbol{w}^*_C\right\|$ monotonically increases) by increasing C, towards its global minimum. If the minimum has been reached, instead, the value $\left\|\boldsymbol{w}^*_C\right\|$ does not change even if C is increased. Refer to Fig. 2 for a graphical representation of the results.

The fundamental result of Theorem 6 allows to derive an approach for solving the Ivanov and Morozov formulations, by exploiting solvers designed for Tikhonov problems.

Concerning the Ivanov regularization formulation, the approach is graphically presented in Fig. 3 and detailed in Algorithm 1. A starting value for the regularization parameter C_{start} is defined: if the solution computed with the Tikhonov formulation is such that $\left\|\boldsymbol{w}^*_{C_{\text{start}}}\right\| > w_{\text{MAX}}$, then the optimal solution for the Ivanov formulation lies on the boundary of the hypothesis space. Since the optimal solution $\left\|\boldsymbol{w}^*_C\right\|$ of Tikhonov Problem (4) monotonically increases (and then $\hat{L}^*_C$ monotonically decreases) by increasing C, it is sufficient to search for $\left\|\boldsymbol{w}^*_C\right\| = w_{\text{MAX}}$ and obtain the solution $\boldsymbol{w}^*_{w_{\text{MAX}}}$ and $\hat{L}^*_{w_{\text{MAX}}}$. For this purpose, a simple bisection algorithm can be exploited. If, instead, $\left\|\boldsymbol{w}^*_{C_{\text{start}}}\right\| \leq w_{\text{MAX}}$, it is possible to search for $C = C_\infty$ such that the solution to the minimization problem does not change even if we keep increasing C, as proven in Theorem 6. If $\left\|\boldsymbol{w}^*_C\right\| < w_{\text{MAX}}$ and the value of $\left\|\boldsymbol{w}^*_C\right\|$ does not change with C, the solution has been found; otherwise, the value of C for which $\left\|\boldsymbol{w}^*_C\right\| = w_{\text{MAX}}$ shall be identified through a bisection procedure. Note that this technique allows to find the smoothest feasible solution to the Ivanov problem, accordingly to the further constraint added when introducing the Ivanov formulation.

Concerning the Morozov regularization formulation, the approach is graphically presented in Fig. 4 and detailed in Algorithm 2. Analogously to the algorithm for Ivanov problems, the approach is initialized with the value C_{start}. If the corresponding empirical error is larger than

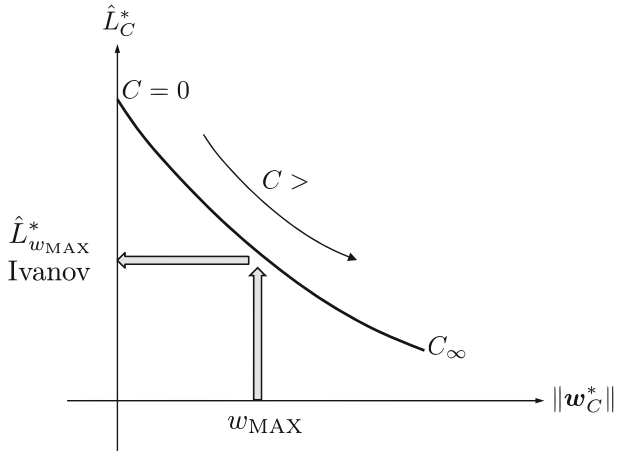

Fig. 3 Graphical representation of the algorithm for solving the Ivanov Problem (7) by exploiting the equivalent Tikhonov formulation of Problem (4)

the threshold $\hat{L}_{\mathrm{MAX}}$, C is increased until it reaches the boundary of the hypothesis space[2] (i.e. $\hat{L}_C^* = \hat{L}_{\mathrm{MAX}}$), and a bisection algorithm can be exploited for this purpose. If the empirical error is smaller than the threshold, C must be decreased in order to identify C_∞, i.e. the value for which $\|\boldsymbol{w}_C^*\|$ starts varying as C decreases. However, we recommend to perform a preliminary step, which consists in checking the error performed by a degenerate model $\boldsymbol{w} = \mathbf{0}$, i.e. the smoothest possible model: if the empirical error for this degenerate model is below the threshold $\hat{L}_{\mathrm{MAX}}$, then this is the solution to the Morozov regularization problem.

There is an underlying assumption in order to be sure that the algorithms will find the solution, namely that the empirical risk has a minimizer.

As a final remark, note that, in principle, one can compute the entire regularization path (Hastie et al. 2004; Gunter and Zhu 2005; Bach et al. 2005; Friedman et al. 2010; Park and Hastie 2007) (i.e. the solution for all values of C or λ) in place of solving several different problems, formulated in accordance with Tikhonov regularization (Algorithms 1 and 2): the final solution is chosen as the one that satisfies the properties discussed above. This approach is generally not convenient, from a computational point of view, when solving one Ivanov (or Morozov) problem for a particular configuration of w_{MAX} (or $\hat{L}_{\mathrm{MAX}}$). However, when coping with model selection and/or error estimation in the SRM framework, several values of the hyperparameters must be explored: as a consequence, in this scenario, computing the entire regularization path can be beneficial from the computational perspective.

5 Implications on learning

We show in this section that the Tikhonov formulation is not directly linked to the SRM-based learning: it is only used as a workaround in order to exploit its computational advantage over the Ivanov formulation (Vapnik 1998). In fact, the SRM framework always resorts to the Ivanov approach, even if implicitly. Other learning frameworks, instead, like the Algorithmic

[2] Note that, in principle, no values of C could exist which allow to find a solution performing better than the pre-determined threshold. However, we neglect this case by hypothesis in our analysis.

Algorithm 1 Algorithm for solving the Ivanov Problem (7) by exploiting the equivalent Tikhonov formulation of Problem (4)

Require: $(z_1, \ldots, z_n)$, $K(\cdot,\cdot)$, w_{MAX} and the tolerance τ

$C = C_{\mathrm{start}}$
$\left(\left\|\boldsymbol{w}_C^*\right\|, \hat{L}_C^*\right) =$ Eq. (4)
if $\left\|\boldsymbol{w}_C^*\right\| > w_{\mathrm{MAX}}$ **then**
  while $\left\|\boldsymbol{w}_C^*\right\| > w_{\mathrm{MAX}}$ **do**
    $C_{\mathrm{up}} = C$
    $C = \frac{C}{2}$
    $\left(\left\|\boldsymbol{w}_C^*\right\|, \hat{L}_C^*\right) =$ Eq. (4)
    if $\left\|\boldsymbol{w}_C^*\right\| < w_{\mathrm{MAX}}$ **then**
      $C_{\mathrm{low}} = C$
      break
    end if
  end while
else
  $w_{\mathrm{new}} = \infty$
  while $\left\|\boldsymbol{w}_C^*\right\| < w_{\mathrm{MAX}}$ **do**
    $C_{\mathrm{low}} = C$
    $C = C * 2$
    $w_{\mathrm{old}} = w_{\mathrm{new}}$
    $\left(\left\|\boldsymbol{w}_C^*\right\|, \hat{L}_C^*\right) =$ Eq. (4)
    if $\left\|\boldsymbol{w}_C^*\right\| > w_{\mathrm{MAX}}$ **then**
      $C_{\mathrm{up}} = C$
      break
    end if
    $w_{\mathrm{new}} = \left\|\boldsymbol{w}_C^*\right\|$
    if $\left|\frac{\|w_{\mathrm{new}} - w_{\mathrm{old}}\|}{\|w_{\mathrm{new}} + w_{\mathrm{old}}\|}\right| < \tau$ **then**
      return $\boldsymbol{w}_C^*$
    end if
  end while
end if
while $\left|\frac{\left\|\boldsymbol{w}_C^*\right\| - w_{\mathrm{MAX}}}{\left\|\boldsymbol{w}_C^*\right\| + w_{\mathrm{MAX}}}\right| > \tau$ **do**
  $C = C_{\mathrm{low}} + \frac{C_{\mathrm{up}} - C_{\mathrm{low}}}{2}$
  $\left(\left\|\boldsymbol{w}_C^*\right\|, \hat{L}_C^*\right) =$ Eq. (4)
  if $\left\|\boldsymbol{w}_C^*\right\| > w_{\mathrm{MAX}}$ **then**
    $C_{\mathrm{up}} = C$
  end if
  if $\left\|\boldsymbol{w}_C^*\right\| < w_{\mathrm{MAX}}$ **then**
    $C_{\mathrm{low}} = C$
  end if
end while
return $\boldsymbol{w}_C^*$

Stability (Bousquet and Elisseeff 2002), can be linked to both formulations, therefore enabling the exploitation of the approach with the most favorable learning characteristics.

For this purpose, let us consider a 1-bounded l-Lipschitz loss function:

$$\ell(h(\boldsymbol{x}), y) \in [0, 1], \quad |\ell(h_1(\boldsymbol{x}), y) - \ell(h_2(\boldsymbol{x}), y)| \leq l|h_1(\boldsymbol{x}) - h_2(\boldsymbol{x})|, \tag{62}$$

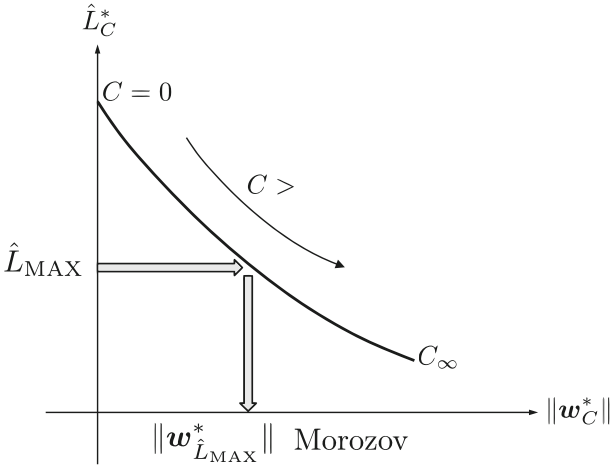

Fig. 4 Graphical representation of the algorithm for solving the Morozov Problem (10) by exploiting the equivalent Tikhonov formulation of Problem (4)

and a kernel class such that $K(\boldsymbol{x}, \boldsymbol{x}) \leq k^2$, where $k > 0$ is a constant. In this case, the generalization error of a classifier can be bounded through the Expected Rademacher Complexity $R(\mathcal{H})$ (Koltchinskii 2001; Bartlett and Mendelson 2003; Anguita et al. 2012). In particular, the latter is an upper bound of the expected difference between $L(h)$ and $\hat{L}(h)$ (Koltchinskii 2001; Bartlett and Mendelson 2003; Anguita et al. 2012), i.e. the Expected Uniform Deviation $U(h)$:

$$U(h) = \mathbb{E}_{\mathcal{D}_n} \sup_{h \in \mathcal{H}} \left[L(h) - \hat{L}(h) \right] \leq \mathbb{E}_{\mathcal{D}_n} \mathbb{E}_\sigma \sup_{h \in \mathcal{H}} \frac{2l}{n} \sum_{i=1}^{n} \sigma_i h(\boldsymbol{x}) = 2l R(\mathcal{H}). \tag{63}$$

Since the Rademacher Complexity $\hat{R}(\mathcal{H}) = \mathbb{E}_\sigma \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i h(\boldsymbol{x})$ and the Uniform Deviation $\hat{U}(h) = \sup_{h \in \mathcal{H}} [L(h) - \hat{L}(h)]$ are bounded difference functions (McDiarmid 1989; Bartlett and Mendelson 2003), the following bound holds with probability $(1 - \delta)$ (Bartlett and Mendelson 2003):

$$L(h) \leq \hat{L}(h) + 2l \hat{R}(\mathcal{H}) + 3 \sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2n}} \tag{64}$$

In order to compute $\hat{R}(\mathcal{H})$, according to SRM, we should fix $\mathcal{H}$ and, for this purpose, an Ivanov regularization approach is the natural choice. The Tikhonov and Morozov approaches, instead, cannot be exploited for computing $\hat{R}(\mathcal{H})$, since $\mathcal{H}$ varies depending on the observations: this problem is also underlined and addressed in Shawe-Taylor et al. (1998). The computation of $\hat{R}(\mathcal{H})$ can be performed by solving a Tikhonov (or Morozov) problem and, then, by using the obtained solution as the constraint in the Ivanov formulation (Anguita et al. 2012). In other words, the Ivanov formulation cannot be avoided.

A possible workaround could consist in bounding $\hat{R}(\mathcal{H})$ by noting that, for kernel classes, $b = 0$, and $\|\boldsymbol{w}\| \leq w_{\mathrm{MAX}}$ the following inequality holds (Bartlett and Mendelson 2003; Bousquet and Elisseeff 2002; Poggio et al. 2002):

Algorithm 2 Algorithm for solving the Morozov Problem (10) by exploiting the equivalent Tikhonov formulation of Problem (4)

Require: $(z_1, \ldots, z_n)$, $K(\cdot,\cdot)$, $\hat{L}_{\mathrm{MAX}}$ and the tolerance τ

if $L_0 = \hat{L}(\boldsymbol{w} = \mathbf{0}) < \hat{L}_{\mathrm{MAX}}$ **then**
  return $\boldsymbol{w}^*_{C=0} = \mathbf{0}$
end if
$C = C_{\mathrm{start}}$
$\left(\|\boldsymbol{w}^*_C\|, \hat{L}^*_C\right) =$ Eq. (4)
if $\hat{L}^*_C < \hat{L}_{\mathrm{MAX}}$ **then**
  while $\hat{L}^*_C < \hat{L}_{\mathrm{MAX}}$ **do**
    $C_{\mathrm{up}} = C$
    $C = \frac{C}{2}$
    $\left(\|\boldsymbol{w}^*_C\|, \hat{L}^*_C\right) =$ Eq. (4)
    if $\hat{L}^*_C > \hat{L}_{\mathrm{MAX}}$ **then**
      $C_{\mathrm{low}} = C$
      break
    end if
  end while
else
  $L_{\mathrm{new}} = \infty$
  while $\hat{L}^*_C > \hat{L}_{\mathrm{MAX}}$ **do**
    $C_{\mathrm{low}} = C$
    $C = C * 2$
    $L_{\mathrm{old}} = L_{\mathrm{new}}$
    $\left(\|\boldsymbol{w}^*_C\|, \hat{L}^*_C\right) =$ Eq. (4)
    if $\hat{L}^*_C < \hat{L}_{\mathrm{MAX}}$ **then**
      $C_{\mathrm{up}} = C$
      break
    end if
    $L_{\mathrm{new}} = \hat{L}^*_C$
    if $\left|\frac{L_{\mathrm{new}} - L_{\mathrm{old}}}{L_{\mathrm{new}} + L_{\mathrm{old}}}\right| < \tau$ **then**
      return $\nexists$ Solution
    end if
  end while
end if
while $\left|\frac{\hat{L}^*_C - \hat{L}_{\mathrm{MAX}}}{\hat{L}^*_C + \hat{L}_{\mathrm{MAX}}}\right| > \tau$ **do**
  $C = C_{\mathrm{low}} + \frac{C_{\mathrm{up}} - C_{\mathrm{low}}}{2}$
  $\left(\|\boldsymbol{w}^*_C\|, \hat{L}^*_C\right) =$ Eq. (4)
  if $\hat{L}^*_C < L_{\mathrm{MAX}}$ **then**
    $C_{\mathrm{up}} = C$
  end if
  if $\hat{L}^*_C > L_{\mathrm{MAX}}$ **then**
    $C_{\mathrm{low}} = C$
  end if
end while
return $\boldsymbol{w}^*_C$

$$\frac{w_{\text{MAX}}}{\sqrt{2n}}\sqrt{\sum_{i=1}^{n} K(\boldsymbol{x}_i,\boldsymbol{x}_i)} \leq \hat{R}(\mathcal{H}) \leq \frac{w_{\text{MAX}}}{n}\sqrt{\sum_{i=1}^{n} K(\boldsymbol{x}_i,\boldsymbol{x}_i)} \leq \frac{k\, w_{\text{MAX}}}{\sqrt{n}}. \tag{65}$$

The proof of the upper bound is straightforward (Bartlett and Mendelson 2003), while the proof of the lower bound is based on the Khintchine inequality (Haagerup 1981). Note that the result of Eq. (65) shows exactly the same results achieved in Bartlett (1998), where Vapnik's approach is exploited (Vapnik 1998): the size of the weights of $\boldsymbol{w}$ is more important than the dimensionality of $\boldsymbol{w}$.

By using the above inequality, the learning process can be performed using the Tikhonov (or Morozov) formulation, without resorting to the Ivanov one. In fact, it is sufficient to set $w_{\text{MAX}} = \|\boldsymbol{w}^*\|$, where $\boldsymbol{w}^*$ is the solution of the Tikhonov (or Morozov) problem. Note that a further indication of the direct link between SRM and the Ivanov formulation is that, to the best knowledge of the authors, there is no explicit upper or lower bound of $\hat{R}(\mathcal{H})$ as a function of C (or $\hat{L}_{\text{MAX}}$).

The same conclusion can be drawn by exploiting the Local version of the Rademacher Complexity (Koltchinskii 2006; Bartlett et al. 2005) for which an equivalent version of Eq. (65) exists (Cortes et al. 2013; Bartlett et al. 2005; Mendelson 2003).

We can link, instead, both the Ivanov and Tikhonov formulations to the Algorithmic Stability framework (Bousquet and Elisseeff 2002; Poggio et al. 2004; Tomasi 2004; Oneto et al. 2014; Elisseeff et al. 2005; Evgeniou et al. 2004). Bousquet and Elisseeff (2002) showed that, under some mild conditions, it is possible to bound the generalization error of a learning algorithm without aprioristically defining a set of hypothesis: in particular, we consider the notion of Uniform Stability which provides the sharpest bounds. Let us define the Uniform Stability as the constant β such that:

$$\forall \mathcal{D}_n, (\boldsymbol{x}, y): \quad |\ell(h_1(\boldsymbol{x}), y) - \ell(h_2(\boldsymbol{x}), y)| \leq \beta \tag{66}$$

where h_1 is learned using the whole set $\mathcal{D}_n$, while h_2 is trained on the set, obtained by removing one sample from $\mathcal{D}_n$. Then, the following bound can be derived:

$$L(h) \leq \hat{L}(h) + 2\beta + (4n\beta + 1)\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2n}} \tag{67}$$

The Uniform Stability β can be upper bounded when the Tikhonov formulation is used for learning (Bousquet and Elisseeff 2002), in fact:

$$|\ell(h_1(\boldsymbol{x}), y) - \ell(h_2(\boldsymbol{x}), y)| \leq l|h_1(\boldsymbol{x}) - h_2(\boldsymbol{x})| \leq l^2 k^2 C \tag{68}$$

Moreover, a similar bound can be found, when using the Ivanov regularization scheme:

$$|\ell(h_1(\boldsymbol{x}), y) - \ell(h_2(\boldsymbol{x}), y)| \leq l|h_1(\boldsymbol{x}) - h_2(\boldsymbol{x})| \leq 2lkw_{\text{MAX}} \tag{69}$$

Therefore, in the Algorithmic Stability framework, both C and w_{MAX} can be used for directly controlling the generalization ability of a learning algorithm.

However, it is worthwhile noting that, given a particular training set, even if we were able to find C and w_{MAX}, such that the solutions to the Tikhonov and Ivanov formulations are equivalent, the stability of the two learning procedures could be different. In other words, let us consider a training set $\mathcal{D}_n$ and let us solve the Tikhonov formulation for a particular value of C, then the solution will be $\boldsymbol{w}_C^*$. If we set $w_{\text{MAX}} = \left\|\boldsymbol{w}_C^*\right\|$ and solve the Ivanov formulation, then the same solution of the Tikhonov one will be found, given the results of Sect. 4. This means that the Tikhonov and the Ivanov formulations are apparently indistinguishable,

since they give the same result when applied to the same data. Moreover, if we analyze the two procedures in the usual SRM framework, we can conclude that the two learning machines choose functions from the same set and therefore have the same associated risk. The Algorithmic Stability, instead, gives more insight into the learning process. In fact, even if the two procedures give the same solution, the risk associated to them can be different. This reflects the fact that the two procedures are, indeed, different despite choosing the same final model. The Tikhonov formulation, for a given C, chooses from a set of hypothesis which depends on $\mathcal{D}_n$, the Ivanov formulation, instead, will always choose from the same set of hypothesis, for a given w_{MAX}, regardless of the available $\mathcal{D}_n$. Consequently, given a specific problem, it would be possible to adopt the formulation characterized by the best learning properties.

The Algorithmic Stability opens also another perspective over the Tikhonov and the Ivanov formulations. Given the solution of the Tikhonov formulation for a given C, one can adopt the Algorithmic Stability bound of Eq. (69) by taking $w_{\text{MAX}} = \left\| \boldsymbol{w}_C^* \right\|$, since the solution of the two formulations would be the same in this setting. In practice, we are pretending to use the Ivanov formulation, since this is more advantageous when the bound of Eq. (69) is tighter. Consequently, the two formulations may have different learning capabilities for different datasets, and the bounds of Eqs. (68) and (69) can tell us which is the best one for the problem under exam: it is only necessary to choose the formulation with the smaller associated risk.

In order to better clarify the above analysis, we show here, through a simple example, that the class $\mathcal{H}$ varies based on the observations in the case of the Tikhonov formulation. Let us consider a one dimensional linear classification problem, where $h(x) = w \cdot x$, and the following datasets:

- TOY-1—n samples in $x = -1$ with $y = -1$; n samples in $x = 1$ with $y = 1$; one sample in $x = 1 + \Delta$, with $y = 1$ (Fig. 5).
- TOY-2—n samples in $x = -1$ with $y = -1$; n samples in $x = 1$ with $y = 1$; one sample in $x = 1 + \Delta$, with $y = -1$ (Fig. 6).

Figure 7 shows the trend of $|w^*|$ as C is varied in $C \in [10^{-6}, 10^4]$, where w^* is the solution of the Tikhonov formulation, when $n = 5$, $\Delta = 8$, and the Hinge loss function (Vapnik 1998) is used. It clearly emerges that the class of functions $\mathcal{H}$ is not determined by C: in fact, for a given value of C, the solution w^* changes depending on the training set.

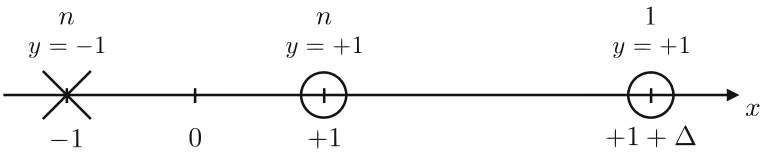

Fig. 5 TOY-1

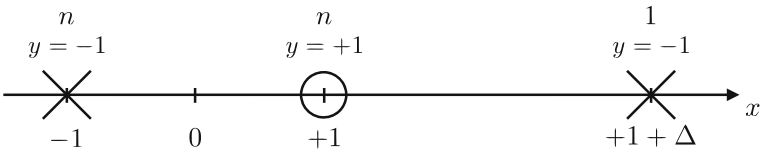

Fig. 6 TOY-2

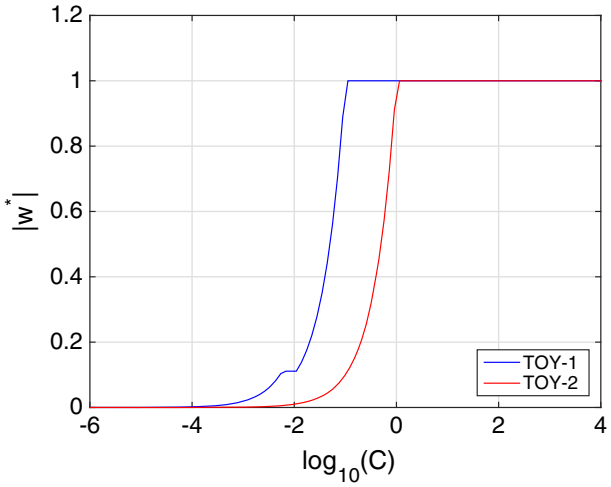

Fig. 7 Trend of $|w^*|$, where w^* is the solution of the Tikhonov formulation, for TOY-1 and TOY-2

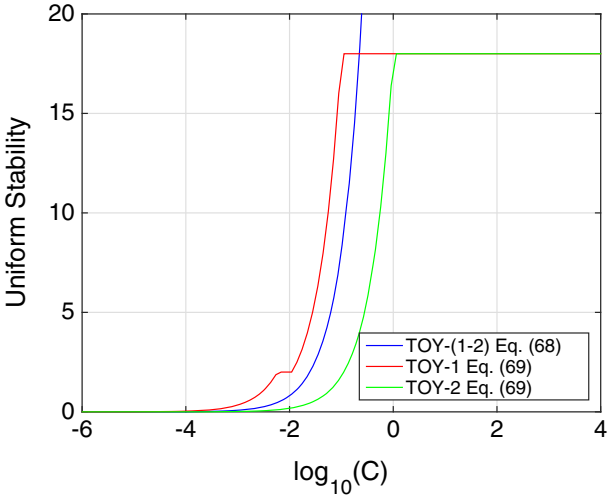

Fig. 8 Uniform Stability by using Eqs. (68) and (69)

The same toy problems can be exploited to show that, using the notion of Uniform Stability, either the Tikhonov or the Ivanov formulations can be adopted, based on the one that shows the best learning properties. Figure 8 shows the value of the Uniform Stability by using Eqs. (68) and (69) as C is varied, where the same experimental setup as above is applied (note that $l = 1$ when the Hinge loss function is exploited). Obviously, if the bound of Eq. (68) is used, the Uniform Stability is the same for the two problems. Instead, if we use the bound of Eq. (69), the Uniform Stability of TOY-1 is larger respect to the one computed with Eq. (68), while the Uniform Stability of TOY-2 is smaller.

As a final remark, it is worth noting that, to the best of our knowledge, a link between the Morozov formulation and the Uniform Stability has not been found yet.

6 Tikhonov, Ivanov and Morozov formulations for support vector machine classifiers learning

The approches proposed in the previous Section are quite general and can be applied to any convex loss function. We will focus here on the SVM classifier, instead, and its hinge-loss function and show that the procedures described by Algorithms 1 and 2 can be further refined and improved.

The hinge loss function is defined as:

$$\ell\left(h(\boldsymbol{x}), y\right) = \xi = \max\left[0, 1 - yh(\boldsymbol{x})\right] = \left|1 - yh(\boldsymbol{x})\right|_+, \tag{70}$$

where $|\cdot|_+ = \max(0, \cdot)$. Then, the T-SVM learning problem can be formulated as follows:

$$\begin{aligned} h: \quad \arg\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \quad & \sum_{i=1}^{n} \xi_i + \frac{\lambda}{2}\|\boldsymbol{w}\|^2 \\ & y_i\left(\boldsymbol{w}\cdot\boldsymbol{\phi}(\boldsymbol{x}_i) + b\right) \geq 1 - \xi_i \\ & \xi_i \geq 0, \qquad i \in \{1, \ldots, n\}, \end{aligned} \tag{71}$$

or equivalently:

$$\begin{aligned} h: \quad \arg\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \quad & \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n} \xi_i \\ & y_i\left(\boldsymbol{w}\cdot\boldsymbol{\phi}(\boldsymbol{x}_i) + b\right) \geq 1 - \xi_i \\ & \xi_i \geq 0, \qquad i \in \{1, \ldots, n\}. \end{aligned} \tag{72}$$

The I-SVM formulation is:

$$\begin{aligned} h: \quad \arg\min_{\boldsymbol{w},b,\boldsymbol{\xi},\|\boldsymbol{w}\|} \quad & \sum_{i=1}^{n} \xi_i \\ & \|\boldsymbol{w}\|^2 \leq w_{\text{MAX}}^2, \\ & y_i\left(\boldsymbol{w}\cdot\boldsymbol{\phi}(\boldsymbol{x}_i) + b\right) \geq 1 - \xi_i \\ & \xi_i \geq 0, \qquad i \in \{1, \ldots, n\}, \end{aligned} \tag{73}$$

Finally, the M-SVM formulation is:

$$\begin{aligned} h: \quad \arg\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \quad & \frac{1}{2}\|\boldsymbol{w}\|^2 \\ & \sum_{i=1}^{n} \xi_i \leq \hat{L}_{\text{MAX}}, \\ & y_i\left(\boldsymbol{w}\cdot\boldsymbol{\phi}(\boldsymbol{x}_i) + b\right) \geq 1 - \xi_i \\ & \xi_i \geq 0, \qquad i \in \{1, \ldots, n\}. \end{aligned} \tag{74}$$

6.1 Training a classifier with T-SVM

The most common approach for training a T-SVM classifier relies on solving the dual formulation of Problem (71):

$$\min_{\boldsymbol{\alpha}} \quad \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j K\left(\boldsymbol{x}_i,\boldsymbol{x}_j\right) - \sum_{i=1}^{n}\alpha_i$$
$$\sum_{i=1}^{n}\alpha_i y_i = 0$$
$$0 \leq \alpha_i \leq C, \qquad i \in \{1,\ldots,n\}, \tag{75}$$

where the classification function is reformulated accordingly:

$$h\left(\boldsymbol{x}\right) = \sum_{i=1}^{n}\alpha_i y_i K\left(\boldsymbol{x}_i,\boldsymbol{x}\right) + b. \tag{76}$$

One of the most well-known approaches for solving Problem (75) is the Sequential Minimal Optimization (SMO) algorithm (Platt 1998; Keerthi et al. 2001; Keerthi and Gilbert 2002; Fan et al. 2005), though several alternatives are available in literature.[3]

6.2 Training a classifier with I-SVM

The first alternative for effectively solving I-SVM consists in exploiting Algorithm 1, where the T-SVM Problem (71) (or (75)) is used in place of Problem (4) and, analogously, I-SVM Problem (73) replaces Problem (7).

A second technique, tailored for I-SVM learning, consists in using the results of Hastie et al. (2004) in order to reduce the computational burden of Algorithm 1. In their work, Hastie et al. (2004) prove that the solution path of T-SVM, obtained by varying C, is piecewise linear. This means that the optimal solution $\left(\boldsymbol{\alpha}_T^*, b_T^*\right)$ linearly changes as C is increased, with the partial exception of S change-points, which can be effectively identified, as the computational cost to find them is comparable to the burden required to find one solution to Problem (75) (Hastie et al. 2004). Moreover, the authors also proved that $\exists C_\infty < \infty$ such that $\forall C \geq C_\infty$ the solution $\left(\boldsymbol{w}_T^*, b_T^*\right)$ does not vary with C, i.e. a particular case of the more general Theorem 6, presented in Sect. 4. Moreover, note that, when SVM is concerned, the equivalence between I-SVM, T-SVM and M-SVM holds not only for the norm of the solutions, but for the solution itself (Hastie et al. 2004). Consequently, we can exploit the information embedded into the regularization path to improve Algorithm 1, when addressing I-SVM learning.

Let $RP_{[33]}$ be the implementation of the method, proposed in Hastie et al. (2004), allowing to identify the solution path for the T-SVM Problem (75):

$$\left\{\left(\boldsymbol{\alpha}_T^*(C_0), b_T^*(C_0), C_0\right), \ldots, \left(\boldsymbol{\alpha}_T^*(C_S), b_T^*(C_S), C_S\right)\right\} = RP_{[33]}, \tag{77}$$

where, in particular, given Theorem 6 and the results of Hastie et al. (2004), $C_S = C_\infty$. As the solution varies linearly, with respect to C, in the span between two consecutive change-points C_{i-1} and C_i, with $i \in \{1,\ldots,S\}$, then:

$$\boldsymbol{\alpha}_T^*(C_i) = \boldsymbol{m}_T^*(C_i)C + \boldsymbol{o}_T^*(C_i), \quad \text{for} \quad C \in \left[C_{i-1}, C_i\right] \tag{78}$$
$$\boldsymbol{m}_T^*(C_i) = \frac{\boldsymbol{\alpha}_T^*(C_i) - \boldsymbol{\alpha}_T^*(C_{i-1})}{C_i - C_{i-1}}, \quad \boldsymbol{o}_T^*(C_i) = \boldsymbol{m}_T^*(C_i)C_{i-1} + \boldsymbol{\alpha}_T^*(C_{i-1}). \tag{79}$$

[3] Refer to, for example, the survey in Shawe-Taylor and Sun (2011).

We can also compute $\left\|\boldsymbol{w}_C^*\right\|^2$ for $C \in \left[C_{i-1}, C_i\right]$:

$$\begin{aligned}\left\|\boldsymbol{w}_C^*\right\|^2 &= \sum_{i=1}^{n}\sum_{i=1}^{n}\left[\left(\boldsymbol{m}_T^*(C_i)\right)_i C + \left(\boldsymbol{o}_T^*(C_i)\right)_i\right] \\ &\quad \left[\left(\boldsymbol{m}_T^*(C_i)\right)_j C + \left(\boldsymbol{o}_T^*(C_i)\right)_j\right] y_i y_j K\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) \qquad (80)\\ &= C^2 \sum_{i=1}^{n}\sum_{i=1}^{n}\left(\boldsymbol{m}_T^*(C_i)\right)_i \left(\boldsymbol{m}_T^*(C_i)\right)_j y_i y_j K\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) \\ &\quad + 2C \sum_{i=1}^{n}\sum_{i=1}^{n}\left(\boldsymbol{m}_T^*(C_i)\right)_i \left(\boldsymbol{o}_T^*(C_i)\right)_j y_i y_j K\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) \\ &\quad + \sum_{i=1}^{n}\sum_{i=1}^{n}\left(\boldsymbol{o}_T^*(C_i)\right)_i \left(\boldsymbol{o}_T^*(C_i)\right)_j y_i y_j K\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) \qquad (81)\\ &= P2_T^*(C_i)C^2 + P1_T^*(C_i)C + P0_T^*(C_i) \quad C \in \left[C_{i-1}, C_i\right] \qquad (82)\end{aligned}$$

Given these results we can reformulate Eq. (77) as follows:

$$\begin{aligned}&\left\{\left[P2_T^*(C_1), P2_T^*(C_1), P2_T^*(C_1), \quad C \in [C_0 = 0, C_1]\right],\right. \\ &\left[P2_T^*(C_2), P2_T^*(C_2), P2_T^*(C_2), \quad C \in [C_1, C_2]\right], \\ &\quad \ldots, \\ &\left.\left[P2_T^*(C_S), P2_T^*(C_S), P2_T^*(C), \quad C \in \left[C_{S-1}, C_S = C_\infty\right]\right]\right\} \qquad (83)\\ &= \left\{P2_T^*(C), P2_T^*(C), P2_T^*(C), \quad C \in [0, C_\infty]\right\} \qquad (84)\\ &= \left\{\left\|\boldsymbol{w}^*\right\|(C), \boldsymbol{w}^*(C), \quad C \in [0, C_\infty]\right\} = RP_{[33]}. \qquad (85)\end{aligned}$$

As a consequence, we can reformulate Algorithm 1 by exploiting the results of Hastie et al. (2004): the resulting approach is presented in Algorithm 3.

Algorithm 3 Algorithm for solving I-SVM Problem (73) by exploiting Algorithm 1 and the results of Hastie et al. (2004)

Require: $(z_1, \ldots, z_n)$, $K(\cdot,\cdot)$, w_{MAX} and the tolerance τ

$\left\{\left\|\boldsymbol{w}^*\right\|(C), \boldsymbol{w}^*(C), \quad C \in [0, C_\infty]\right\} = RP_{\text{(Hastie et al. 2004)}}$
if $\left\|\boldsymbol{w}^*\right\|(C_\infty) < w_{\text{MAX}}$ **then**
  return $\boldsymbol{w}^*(C_\infty)$
end if
$C_{\text{up}} = C_\infty$
$C_{\text{low}} = 0$
while $\left|\left\|\boldsymbol{w}^*\right\|(C) - w_{\text{MAX}}\right| > \tau$ **do**
  $C = C_{\text{low}} + \frac{C_{\text{up}} - C_{\text{low}}}{2}$
  if $\left\|\boldsymbol{w}^*\right\|(C) > w_{\text{MAX}}$ **then**
    $C_{\text{up}} = C$
  end if
  if $\left\|\boldsymbol{w}_C^*\right\|(C) < w_{\text{MAX}}$ **then**
    $C_{\text{low}} = C$
  end if
end while
return $\boldsymbol{w}^*(C)$

A third approach is based on the ideas of Martein and Schaible (1987) and Anguita et al. (2010). It exploits, in turn, conventional Linear (LP) and Quadratic Programming (QP) optimization algorithms, as shown in Algorithm 4. The first step consists in discarding the quadratic constraint $\|\boldsymbol{w}\|^2 \leq w_{\text{MAX}}^2$ and using the Representer Theorem to reformulate Problem (73) as follows:

$$\begin{aligned} h: \quad \arg\min_{\boldsymbol{\alpha},b,\boldsymbol{\xi}} \quad & \sum_{i=1}^{n} \xi_i \\ & y_i \left(\sum_{j=1}^{n} \alpha_j y_j K(\boldsymbol{x}_j, \boldsymbol{x}_i) + b \right) \geq 1 - \xi_i \\ & \xi_i \geq 0, \qquad i \in \{1, \ldots, n\} \end{aligned} \tag{86}$$

which is a standard LP problem. After the optimization procedure ends, the value of $\|\boldsymbol{w}\|^2$ is computed and two alternatives arise: if the constraint is satisfied, then the solution found is also the optimal one and the routine ends; otherwise the optimal solution corresponds to $\|\boldsymbol{w}\| = w_{MAX}$ (Boyd and Vandenberghe 2004). In order to find $\boldsymbol{w}$, we have to switch to the dual of Problem (73):

$$\begin{aligned} \mathcal{L}(\boldsymbol{w}, b, \boldsymbol{\xi}, \boldsymbol{\beta}, \gamma, \boldsymbol{\mu}) = & \sum_{i=1}^{n} \xi_i - \frac{\gamma}{2} \left(w_{\text{MAX}}^2 - \|\boldsymbol{w}\|^2 \right) \\ & - \sum_{i=1}^{n} \beta_i \left[y_i \left(\boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}_i) \right) + b \right) - 1 + \xi_i \right] - \sum_{i=1}^{n} \mu_i \xi_i . \end{aligned} \tag{87}$$

Then we can compute the Karush–Kuhn–Tucker (KKT) and the complementary conditions:

$$\frac{\partial \mathcal{L}(\boldsymbol{w}, b, \boldsymbol{\xi}, \boldsymbol{\beta}, \gamma, \boldsymbol{\mu})}{\partial \boldsymbol{w}} = \gamma \boldsymbol{w} - \sum_{i=1}^{n} \beta_i y_i \boldsymbol{\phi}(\boldsymbol{x}_i) = 0 \;\rightarrow\; \boldsymbol{w} = \frac{1}{\gamma} \sum_{i=1}^{n} \beta_i y_i \boldsymbol{\phi}(\boldsymbol{x}_i) \qquad \gamma \neq 0 \tag{88}$$

$$\frac{\partial \mathcal{L}(\boldsymbol{w}, b, \boldsymbol{\xi}, \boldsymbol{\beta}, \gamma, \boldsymbol{\mu})}{\partial b} = -\sum_{i=1}^{n} \beta_i y_i = 0 \tag{89}$$

$$\frac{\partial \mathcal{L}(\boldsymbol{w}, b, \boldsymbol{\xi}, \boldsymbol{\beta}, \gamma, \boldsymbol{\mu})}{\partial \xi_i} = 1 - \beta_i - \mu_i = 0 \;\rightarrow\; \beta_i \leq 1 \tag{90}$$

$$\gamma, \boldsymbol{\beta}, \boldsymbol{\mu}, \xi_i \geq 0 \tag{91}$$

$$\|\boldsymbol{w}\|^2 \leq w_{\text{MAX}}^2 \tag{92}$$

$$y_i \left(\boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}) + b \right) \geq 1 - \xi_i \tag{93}$$

$$\beta_i \left[y_i \left(\boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}) + b \right) - 1 + \xi_i \right] = 0 \tag{94}$$

$$\mu_i \xi_i = 0 \tag{95}$$

$$\gamma \left(w_{\text{MAX}}^2 - \|\boldsymbol{w}\|^2 \right) = 0, \quad \forall i \in \{1, \ldots, n\} \tag{96}$$

The dual formulation is formulated as follows:

$$\begin{aligned}\min_{\boldsymbol{\beta},\gamma} \quad & \frac{1}{2\gamma}\sum_{i=1}^{n}\sum_{j=1}^{n}\beta_i\beta_j y_i y_j K(\boldsymbol{x}_j,\boldsymbol{x}_i) - \sum_{i=1}^{n}\beta_i + \frac{\gamma w_{\text{MAX}}^2}{2} \\ & \sum_{i=1}^{n}\beta_i y_i = 0 \\ & \gamma \geq 0 \\ & 0 \leq \beta_i \leq 1, \qquad i \in \{1,\ldots,n\}. \end{aligned} \tag{97}$$

Note that, if the quadratic constraint $\|\boldsymbol{w}\|^2 \leq w_{\text{MAX}}^2$ were satisfied, γ would equal zero and the dual formulation would not be solvable: this is why, as a first step, we make use of the LP routines for solving Problem (73) and we exploit the dual formulation only if the quadratic constraint is not satisfied. We are interested in solving Problem (97) using conventional QP solvers for SVM (e.g. Platt 1998), therefore we use an iterative optimization technique. The first step consists in fixing the value of $\gamma = \gamma_0 > 0$ and, then, optimizing the cost function with reference to the other dual variables $\boldsymbol{\beta}$. It is easy to see that the term $\frac{\gamma w_{MAX}^2}{2}$ is constant at this stage and can be removed from the expression, so the dual becomes:

$$\begin{aligned}\min_{\boldsymbol{\beta}} \quad & \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\beta_i\beta_j y_i y_j K(\boldsymbol{x}_j,\boldsymbol{x}_i) - \gamma_0\sum_{i=1}^{n}\beta_i \\ & \sum_{i=1}^{n}\beta_i y_i = 0 \\ & 0 \leq \beta_i \leq 1, \qquad i \in \{1,\ldots,n\}. \end{aligned} \tag{98}$$

which is equivalent to the conventional SVM dual problem (75) and can be solved with SMO. The next step consists in updating the value of γ_0. We have to compute the Lagrangian of Problem (97):

$$\begin{aligned}\mathcal{L}(\boldsymbol{\beta},\gamma,\rho,\kappa,\boldsymbol{\eta},\boldsymbol{\nu}) = & \frac{1}{2\gamma}\sum_{i=1}^{n}\sum_{j=1}^{n}\beta_i\beta_j y_i y_j K(\boldsymbol{x}_j,\boldsymbol{x}_i) - \sum_{i=1}^{n}\beta_i + \frac{\gamma w_{\text{MAX}}^2}{2} \\ & - \rho\left(\sum_{i=1}^{n}\beta_i y_i\right) - \kappa\gamma - \sum_{i=1}^{n}\eta_i\beta - \sum_{i=1}^{n}\nu_i(C-\beta_i) \end{aligned} \tag{99}$$

The following derivative of $\mathcal{L}(\boldsymbol{\beta},\gamma,\rho,\kappa,\boldsymbol{\eta},\boldsymbol{\nu})$ is the only one of interest for our purposes:

$$\frac{\partial\mathcal{L}(\boldsymbol{\beta},\gamma,\rho,\kappa,\boldsymbol{\eta},\boldsymbol{\nu})}{\partial\gamma} = 0 = -\frac{1}{2\gamma^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\beta_i\beta_j y_i y_j K(\boldsymbol{x}_j,\boldsymbol{x}_i) + \frac{w_{\text{MAX}}^2}{2} - \kappa \tag{100}$$

Since, from the slackness conditions, we have that $\kappa\gamma = 0$ and since, in the cases of interest, $\gamma > 0$, it must be $\kappa = 0$ and we find the following updating rule for γ_0:

$$\gamma_0 = \frac{\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}\beta_i\beta_j y_i y_j K(\boldsymbol{x}_j,\boldsymbol{x}_i)}}{w_{MAX}}. \tag{101}$$

Algorithm 4 Algorithm for solving I-SVM Problem (73) based on the results of Martein and Schaible (1987) and Anguita et al. (2010)

Require: $(z_1, \ldots, z_n)$, $K(\cdot, \cdot)$, w_{MAX} and the tolerance τ

$\{\boldsymbol{w}, b, \boldsymbol{\xi}\}$ = solve LP problem of Eq. (86)
if $\|\boldsymbol{w}\|^2 < w_{\text{MAX}}^2$ **then**
  return $\{\boldsymbol{w}, b\}$
else
  $\gamma_0^{\text{old}} = \infty$
  γ_0 = Eq. (101)
  while $\left|\frac{\gamma_0 - \gamma_0^{\text{old}}}{\gamma_0 + \gamma_0^{\text{old}}}\right| > \tau$ **do**
    $\gamma_0^{\text{old}} = \gamma_0$
    $\{\boldsymbol{w}, b, \boldsymbol{\xi}\}$ = solve QP problem of Eq. (98)
    γ_0 = Eq. (101)
  end while
end if
return $\{\boldsymbol{w}, b\}$

We iteratively proceed in solving the dual of Problem (98) and updating the value of γ_0 until the termination condition is met:

$$\left|\gamma_0 - \gamma_0^{\text{old}}\right| \leq \tau, \tag{102}$$

where τ is a user-defined tolerance.

The main disadvantage of Algorithm 4 is that it requires the use of two different solvers (LP and QP routines). By exploiting the results of Sect. 4, we can avoid this problem: in fact, we showed that the Lagrange multiplier of the quadratic constraint of Eq. (73) (γ) is equivalent to the regularization parameter C in Eq. (71). Consequently, there exists a value of $\gamma > 0$ for which the solution of Problem (73) is equivalent to the one of Problem (71) for $C = C_\infty$. Consequently, only the QP solver is necessary to train a classifier with I-SVM, as shown in Algorithm 4.

Algorithm 5 Algorithm to solve I-SVM Problem (73) by only requiring a QP solver

Require: $(z_1, \ldots, z_n)$, $K(\cdot, \cdot)$, w_{MAX} and the tolerance τ

$\epsilon = \tau$
$\gamma_0 = \infty$
$w_{\text{old}} = \infty$
repeat
  $\gamma_0^{\text{old}} = \gamma_0$
  $\{\boldsymbol{w}, b, \boldsymbol{\xi}\}$ = solve QP problem of Eq. (98)
  $\gamma_0 = \min\left[\epsilon, \frac{\sqrt{\sum_{i=1}^n \sum_{j=1}^n \beta_i \beta_j y_i y_j K(\boldsymbol{x}_j, \boldsymbol{x}_i)}}{w_{MAX}}\right]$
  if $\gamma_0 == \epsilon$ **then**
    $\epsilon = \epsilon/10$
    if $w_{\text{old}} == \sqrt{\boldsymbol{w} \cdot \boldsymbol{w}}$ **then**
      break
    end if
    $w_{\text{old}} = \sqrt{\boldsymbol{w} \cdot \boldsymbol{w}}$
  end if
until $\left|\frac{\gamma_0 - \gamma_0^{\text{old}}}{\gamma_0 + \gamma_0^{\text{old}}}\right| > \tau$
return $\{\boldsymbol{w}, b\}$

6.3 Training a classifier with M-SVM

In order to propose an effective method to solve M-SVM by exploiting the approaches designed for T-SVM, as a first step we can apply the general-purpose approach of Algorithm 2. Obviously, this technique is characterized by the same drawbacks highlighted in Sect. 6.2 for I-SVM.

A second possibility consists, analogously to I-SVM, in exploiting the results of Hastie et al. (2004) to improve the performance of Algorithm 2. The entire regularization path is obtained by:

$$\left\{\boldsymbol{w}^*(C), b^*(C), \boldsymbol{\xi}^*(C), \quad C \in [0, C_\infty]\right\} = RP_{[33]}, \tag{103}$$

where the computational cost is equal to the one of a single optimization step of Problem (71) since all the variables ($\boldsymbol{\alpha}, b, \boldsymbol{\xi}$ etc.) are piecewise linear in C. Then Algorithm 2 can be modified accordingly as shown in Algorithm 6.

Algorithm 6 Algorithm to solve M-SVM Problem (74) by exploiting Algorithm 2 and the results of Hastie et al. (2004)

Require: $(z_1, \ldots, z_n)$, $K(\cdot, \cdot)$, w_{MAX} and the tolerance τ

$\left\{\boldsymbol{w}^*(C), b^*(C), \boldsymbol{\xi}^*(C), \quad C \in [0, C_\infty]\right\} = RP_{\text{Hastie et al. (2004)}}$
if $\sum_{i=0}^{n} \xi^*(0) < \hat{L}_{\text{MAX}}$ **then**
  return $\boldsymbol{w}^*(0)$
end if
if $\sum_{i=0}^{n} \xi^*(C_\infty) > \hat{L}_{\text{MAX}}$ **then**
  return No solution
end if
$C_{\text{up}} = C_\infty$
$C_{\text{low}} = 0$
while $|\sum_{i=0}^{n} \xi^*(C) - \hat{L}_{\text{MAX}}| > \tau$ **do**
  $C = C_{\text{low}} + \frac{C_{\text{up}} - C_{\text{low}}}{2}$
  if $\sum_{i=0}^{n} \xi^*(C) < \hat{L}_{\text{MAX}}$ **then**
    $C_{\text{up}} = C$
  end if
  if $\sum_{i=0}^{n} \xi^*(C) > \hat{L}_{\text{MAX}}$ **then**
    $C_{\text{low}} = C$
  end if
end while
return $\boldsymbol{w}^*(C)$

A third possibility consists in solving Problem (74) through an ad hoc procedure, which however allows to exploit the large amount of work pursued for designing effective solvers for T-SVM. For this purpose, we start by deriving the dual formulation for M-SVM:

$$\begin{aligned}\mathcal{L}(\boldsymbol{w}, b, \boldsymbol{\xi}, \gamma, \boldsymbol{\beta}, \boldsymbol{\mu}) = \frac{1}{2}\|\boldsymbol{w}\|^2 - \gamma\left(\hat{L}_{\text{MAX}} - \sum_{i=1}^{n}\xi_i\right) \\ - \sum_{i=1}^{n}\beta_i\left[y_i\left(\boldsymbol{w}\cdot\boldsymbol{\phi}(\boldsymbol{x}_i) + b\right) - 1 + \xi_i\right] - \sum_{i=1}^{n}\mu_i\xi_i.\end{aligned} \tag{104}$$

Then we can compute its KKT and the complementary conditions:

$$\frac{\partial\mathcal{L}(\boldsymbol{w}, b, \boldsymbol{\xi}, \gamma, \boldsymbol{\beta}, \boldsymbol{\mu})}{\partial\boldsymbol{w}} = \boldsymbol{w} - \sum_{i=1}^{n}\beta_i y_i\boldsymbol{\phi}(\boldsymbol{x}_i) = 0 \;\rightarrow\; \boldsymbol{w} = \sum_{i=1}^{n}\beta_i y_i\boldsymbol{\phi}(\boldsymbol{x}_i) \tag{105}$$

$$\frac{\partial \mathcal{L}(\boldsymbol{w}, b, \boldsymbol{\xi}, \gamma, \boldsymbol{\beta}, \boldsymbol{\mu})}{\partial b} = -\sum_{i=1}^{n} \beta_i y_i = 0 \tag{106}$$

$$\frac{\partial \mathcal{L}(\boldsymbol{w}, b, \boldsymbol{\xi}, \gamma, \boldsymbol{\beta}, \boldsymbol{\mu})}{\partial \xi_i} = \gamma - \beta_i - \mu_i = 0 \;\rightarrow\; \beta_i \leq \gamma \tag{107}$$

$$\gamma, \boldsymbol{\beta}, \boldsymbol{\mu}, \xi_i \geq 0 \tag{108}$$

$$\sum_{i=1}^{n} \xi_i \leq \hat{L}_{\text{MAX}} \tag{109}$$

$$y_i \left(\boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}) + b\right) \geq 1 - \xi_i \tag{110}$$

$$\beta_i \left[y_i \left(\boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}) + b\right) - 1 + \xi_i\right] = 0 \tag{111}$$

$$\mu_i \xi_i = 0 \tag{112}$$

$$\gamma \left(\hat{L}_{\text{MAX}} - \sum_{i=1}^{n} \xi_i\right) = 0, \quad \forall i \in \{1, \ldots, n\} \tag{113}$$

Finally, we get:

$$\begin{aligned} \min_{\boldsymbol{\beta}, \gamma} \quad & CD(\boldsymbol{\beta}, \gamma) = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \beta_i \beta_j y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) - \sum_{i=1}^{n} \beta_i + \gamma \hat{L}_{\text{MAX}} \\ & \sum_{i=1}^{n} \beta_i y_i = 0 \\ & \gamma \geq 0 \\ & 0 \leq \beta_i \leq \gamma, \qquad i \in \{1, \ldots, n\}, \end{aligned} \tag{114}$$

with

$$h(\boldsymbol{x}) = \sum_{i=1}^{n} \beta_i y_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b. \tag{115}$$

Note that, if we fix $\gamma = \gamma_0$, where γ_0 is a constant value, Problem (114) becomes:

$$\begin{aligned} \min_{\boldsymbol{\beta}, \gamma} \quad & \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \beta_i \beta_j y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) - \sum_{i=1}^{n} \beta_i \\ & \sum_{i=1}^{n} \beta_i y_i = 0 \\ & 0 \leq \beta_i \leq \gamma_0, \qquad i \in \{1, \ldots, n\}, \end{aligned} \tag{116}$$

which thus can be solved with the procedures, designed for T-SVM. As the problem is also convex with respect to γ, an iterative procedure, described in Algorithm 7 and based on Flannery et al. (1992), can be used to optimize it. It is worthwhile underlining that the resulting procedure only requires one QP solver, analogously to Algorithm 5.

7 Benchmarking the performance of I-SVM and M-SVM solvers

We exploit in the following a real-world dataset for the purpose of comparing the algorithms presented in the previous sections. We make use of the DaimlerChrysler dataset (Munder and

Algorithm 7 Algorithm to solve M-SVM Problem (74) by only requiring one QP solver

Require: $(z_1, \ldots, z_n)$, $K(\cdot,\cdot)$, w_{MAX} and the tolerance τ

$\gamma_0 = 0$
$(\boldsymbol{\beta}_0, b_0)$ = Eq. (116)(γ_0)
$CD_0 = CD(\boldsymbol{\beta}_0, \gamma_0)$
$\gamma_0 = 10$
while $CD(\boldsymbol{\beta}_0, \gamma_0) < CD_0$ **do**
  $\gamma_0 = \gamma_0 * 10$
  $(\boldsymbol{\beta}_0, b_0)$ = Eq. (116)(γ_0)
end while
$CD_3 = CD(\boldsymbol{\beta}_0, \gamma_0)$
$\gamma_3 = \gamma_0$, $\gamma_0 = 0$
while $\left|\frac{\gamma_3-\gamma_0}{\gamma_3+\gamma_0}\right| > \tau$ **do**
  $\gamma_1 = \gamma_0 + \frac{1}{3}(\gamma_3 - \gamma_0)$, $\gamma_2 = \gamma_0 + \frac{2}{3}(\gamma_3 - \gamma_0)$
  $(\boldsymbol{\beta}_0, b_0)$ = Eq. (116)(γ_1)
  $CD_1 = CD(\boldsymbol{\beta}_0, \gamma_1)$
  $(\boldsymbol{\beta}_0, b_0)$ = Eq. (116)(γ_2)
  $CD_2 = CD(\boldsymbol{\beta}_0, \gamma_2)$
  if $(CD_1 \leq CD_2)$ **then**
    $\gamma_3 = \gamma_2$
    $CD_3 = CD_2$
  else
    $\gamma_0 = \gamma_1$
    $CD_0 = CD_1$
  end if
end while
return $\{\boldsymbol{w}, b\}$

Gavrila 2006), where half of the 9800 images, consisting of $d = 36 \times 18 = 648$ pixels, contains the picture of a pedestrian, while the other half contains only some general background or other objects. In order to derive more statistically relevant results than a single run, we create 100 replicates of the dataset, where the value of the input patterns are left unchanged while a random array of labels is assigned to the samples, thus emulating a conventional setup for model selection and error estimation through the Rademacher Complexity (Bartlett and Mendelson 2003; Koltchinskii 2006; Anguita et al. 2012).

The algorithms are implemented in FORTRAN 90, compiled by exploiting the Intel Visual Fortran Composer XE compiler (2012), and are run on a Microsoft Windows Server 2008 R2 server with 16 GB RAM and mounting two Intel Xeon E5320 1.86 GHz CPUs.

For our experiments, we exploit a Gaussian kernel function (Keerthi and Lin 2003):

$$K\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) = e^{-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2}{2\sigma^2}}, \tag{117}$$

where the Gaussian width parameter σ is estimated by computing the average distance between patterns belonging to the two classes, according to the rule-of-thumb proposed in Milenova et al. (2005). The SVM regularization parameters C, w_{MAX} and $\hat{L}_{\text{MAX}}$ for T-SVM, I-SVM and M-SVM, respectively, are set by exploiting one-shot procedures as well. In particular, for C we adopt the procedure proposed in Milenova et al. (2005), returning an regularization parameter value that we define $C^{[46]}$ for the sake of simplicity. Consequently, w_{MAX} and $\hat{L}_{\text{MAX}}$ are respectively set to $w_{\text{MAX}}^{[46]}$ and $\hat{L}_{\text{MAX}}^{[46]}$, which are computed by solving T-SVM Problem (71) with $C = C^{[46]}$ on the original DaimlerChrysler dataset. Note that the

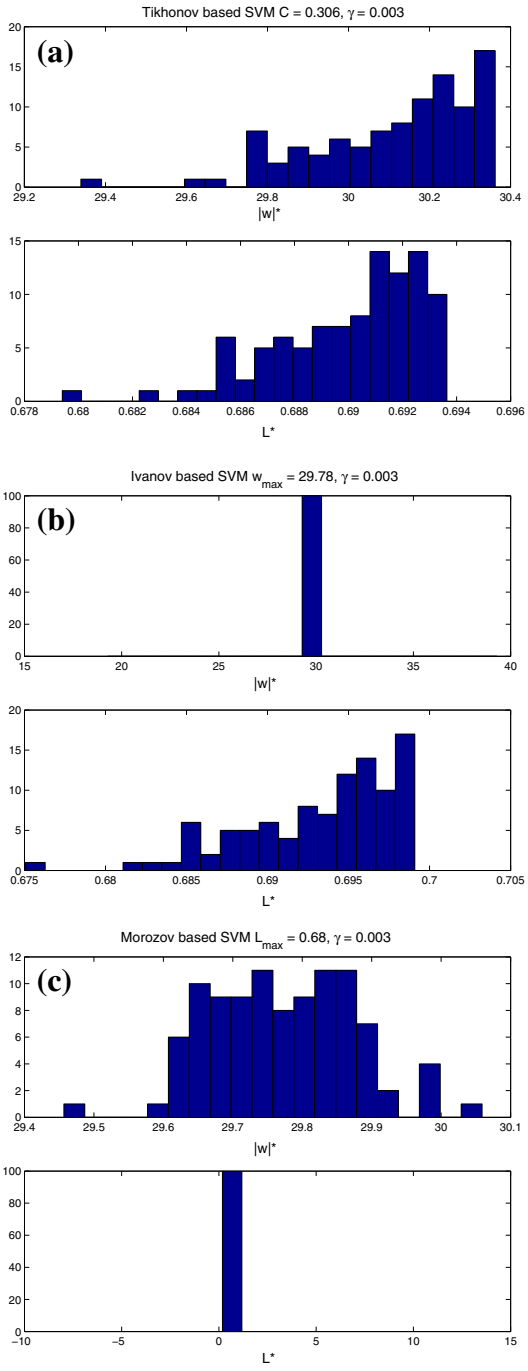

Fig. 9 Distributions of the values $\left(\|\boldsymbol{w}_T^*\|, \hat{L}_T^*\right)$, $\left(\|\boldsymbol{w}_I^*\|, \hat{L}_I^*\right)$ and $\left(\|\boldsymbol{w}_M^*\|, \hat{L}_M^*\right)$ for T-SVM, I-SVM and M-SVM on the datasets used for the experiments. **a** T-SVM Problem (71), **b** I-SVM Problem (73), **c** M-SVM Problem (74)

Table 1 Training time for T-SVM, I-SVM and M-SVM

Regularization approach	Algorithm	Time (min)
T-SVM Problem (71)	Algorithm (Fan et al. 2005)	45.1 ± 3.9
I-SVM Problem (73)	Algorithm 1	75.7 ± 5.4
	Algorithm 3	55.1 ± 4.3
	Algorithm 4	123.7 ± 8.3
	Algorithm 5	47.2 ± 3.7
M-SVM Problem (74)	Algorithm 2	78.7 ± 4.9
	Algorithm 6	52.1 ± 4.5
	Algorithm 7	46.4 ± 3.2

regularization parameter and the parameter of the Gaussian kernel are kept constant for all 100 random replicates.

As highlighted in the previous sections, some solvers are needed in order to derive the solutions to T-SVM, I-SVM and M-SVM. In particular, whenever a QP solver is required (namely in Algorithms 1, 2, 3, 4, 5, 6, and 7), the Sequential Minimal Optimization (SMO) procedure is exploited (Platt 1998; Keerthi et al. 2001); on the contrary, when an LP solver is required (e.g. in Algorithm 4), we exploit the simplex method proposed in Flannery et al. (1992).

Figure 9 shows the distributions of the values $\left(\left\|\boldsymbol{w}_T^*\right\|, \hat{L}_T^*\right)$, $\left(\left\|\boldsymbol{w}_I^*\right\|, \hat{L}_I^*\right)$ and $\left(\left\|\boldsymbol{w}_M^*\right\|, \hat{L}_M^*\right)$ for T-SVM, I-SVM and M-SVM: they are useful to compare the obtained solutions in the three cases. It is worth noting that, when solving a T-SVM learning problem, neither $\left\|\boldsymbol{w}_T^*\right\|$ nor $\hat{L}_T^*$ remain fixed, but they vary depending on the random labels assigned to the samples for computing the Rademacher Complexity of the hypothesis space. This shows the problematic behavior of the T-SVM approach, which does not allow to precisely control the size of the hypothesis space. I-SVM, instead, uses by construction a fixed hypothesis space, while M-SVM shows, as predicted, a fixed empirical error on the training data.

Table 1 shows the average training time, needed by the algorithms detailed in this work for solving T-SVM, I-SVM and M-SVM problems as described above. As expected, T-SVM is the fastest to solve, though Algorithms 5 and 7 are characterized by a comparable performance. Given that the I-SVM shows a clear advantage in terms of hypothesis space control, but only a slight increase in computation time, we believe that the I-SVM should be the preferred approach.

8 Concluding remarks

In this paper, we proved that the three regularization paths for Tikhonov, Ivanov and Morozov regularization are equivalent, provided that mild and easy-to-satisfy conditions hold (such as the convexity of the loss function). In other words, they are the same learning problem seen from three different perspectives.

Traditionally, this reason motivated the exploitation of the Tikhonov approach at the expense of Ivanov and Morozov ones: by leaving unconstrained both the empirical error and the hypothesis space size terms, the Tikhonov formulation is the easiest to solve and several approaches appeared in literature for this purpose in the last decades. However, the capability of fixing one of the two quantities, which control the learning process, is of importance in order

to derive more insights and more refined approaches dealing with performance assessment of learnt models, especially taking in account recent advances and refinements of the SRM principle (Vapnik 1998; Bartlett and Mendelson 2003; Koltchinskii 2006; Bousquet et al. 2004; Shawe-Taylor et al. 1998).

This is particularly true for Ivanov regularization, which represents the most intuitive and direct implementation of the SRM principle, as also underlined by Vapnik in his seminal work (Vapnik 1998). When the SVM was introduced as a Tikhonov formulation, an unmet gap was created between the capacity of effectively and intuitively controlling the hypothesis space size, typical of the Ivanov approach, and the performance in model training, namely the reason-why Tikhonov regularization was chosen. this is the reason that leaded us to study how this gap could be filled: in particular, we proposed effective and easy-to-implement approaches to solve the I-SVM, without neglecting the huge amount of work in the last years dedicated to solving T-SVM.

References

Anguita, D., Ghio, A., Greco, N., Oneto, L., & Ridella, S. (2010). Model selection for support vector machines: Advantages and disadvantages of the machine learning theory. In *International joint conference on neural networks* (pp. 1–8).

Anguita, D., Ghio, A., Oneto, L., & Ridella, S. (2011). The impact of unlabeled patterns in rademacher complexity theory for kernel classifiers. In: *Advances in neural information processing systems*, (pp. 585–593).

Anguita, D., Ghio, A., Oneto, L., & Ridella, S. (2011). In-sample model selection for support vector machines. In *International joint conference on neural networks* (pp. 1154–1161).

Anguita, D., Ghio, A., Oneto, L., & Ridella, S. (2012). In-sample and out-of-sample model selection and error estimation for support vector machines. *IEEE Transactions on Neural Networks and Learning Systems*, *23*(9), 1390–1406.

Anthony, M. (2001). *Discrete mathematics of neural networks: Selected topics*. Philadelphia: Society for Industrial Mathematics.

Aronszajn, N. (1951). *Theory of reproducing kernels*. Cambridge: Harvard University.

Bach, F. R., Thibaux, R., & Jordan, M. I. (2005). Computing regularization paths for learning multiple kernels. In *Advances in neural information processing systems*.

Bartlett, P., Bousquet, O., & Mendelson, S. (2005). Local rademacher complexities. *Annals of Statistics*, *33*(4), 1497–1537.

Bartlett, P., Jordan, M., & McAuliffe, J. (2006). Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, *101*(473), 138–156.

Bartlett, P. L. (1998). The sample complexity of pattern classification with neural networks: The size of the weights is more important than the size of the network. *IEEE Transactions on Information Theory*, *44*(2), 525–536.

Bartlett, P. L., & Mendelson, S. (2003). Rademacher and Gaussian complexities: Risk bounds and structural results. *The Journal of Machine Learning Research*, *3*, 463–482.

Bauschke, H. H., & Combettes, P. L. (2011). *Convex analysis and monotone operator theory in Hilbert spaces*. New York: Springer.

Berlinet, A., & Thomas-Agnan, C. (2004). *Reproducing kernel Hilbert spaces in probability and statistics*. New York: Springer.

Bishop, C. (1995). Training with noise is equivalent to Tikhonov regularization. *Neural Computation*, *7*(1), 108–116.

Bousquet, O., Boucheron, S., & Lugosi, G. (2004). Introduction to statistical learning theory. In *Advanced lectures on machine learning* (pp. 169–207).

Bousquet, O., & Elisseeff, A. (2002). Stability and generalization. *The Journal of Machine Learning Research*, *2*, 499–526.

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge: Cambridge University Press.

Collins, M., Schapire, R., & Singer, Y. (2002). Logistic regression, AdaBoost and Bregman distances. *Machine Learning*, *48*(1–3), 253–285.

Cortes, C., Kloft, M., & Mohri, M. (2013). Learning kernels using local Rademacher complexity. In *Advances in neural information processing systems*.
Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.
Dinuzzo, F., & Schölkopf, B. (2012). The representer theorem for hilbert spaces: A necessary and sufficient condition. Arxiv preprint arXiv:1205.1928.
Duan, K., Keerthi, S., & Poo, A. (2003). Evaluation of simple performance measures for tuning SVM hyperparameters. *Neurocomputing*, *51*, 41–59.
Elisseeff, A., Evgeniou, T., & Pontil, M. (2005). Stability of randomized learning algorithms. *Journal of Machine Learning Research*, *6*, 55–79.
Evgeniou, T., Pontil, M., & Elisseeff, A. (2004). Leave one out error, stability, and generalization of voting combinations of classifiers. *Machine Learning*, *55*(1), 71–97.
Fan, R., Chen, P., & Lin, C. (2005). Working set selection using second order information for training support vector machines. *Journal of Machine Learning Research*, *6*, 1889–1918.
Feldman, V., Guruswami, V., Raghavendra, P., & Wu, Y. (2009). Agnostic learning of monomials by halfspaces is hard. In *Annual IEEE symposium on foundations of computer science* (pp. 385–394).
Flannery, B., Press, W., Teukolsky, S., & Vetterling, W. (1992). *Numerical recipes in C*. New York: Press Syndicate of the University of Cambridge.
Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1.
Goldstein, A. A. (1977). Optimization of Lipschitz continuous functions. *Mathematical Programming*, *13*(1), 14–22.
Gunter, L., & Zhu, J. (2005). Computing the solution path for the regularized support vector regression. In *Neural information processing systems* (pp. 481–488).
Guyon, I., Saffari, A., Dror, G., & Cawley, G. (2010). Model selection: Beyond the Bayesian/frequentist divide. *The Journal of Machine Learning Research*, *11*, 61–87.
Haagerup, U. (1981). The best constants in the Khintchine inequality. *Studia Mathematica*, *70*(3), 231–283.
Hastie, T., Rosset, S., Tibshirani, R., & Zhu, J. (2004). The entire regularization path for the support vector machine. *The Journal of Machine Learning Research*, *5*, 1391–1415.
Intel: Intel Visual Fortran Composer XE. (2012). http://software.intel.com/en-us/articles/intel-compilers/. Accessed: 18/07/2012.
Ivanov, V. (1976). *The theory of approximate methods and their application to the numerical solution of singular integral equations*. New York: Springer.
Keerthi, S., & Gilbert, E. (2002). Convergence of a generalized SMO algorithm for SVM classifier design. *Machine Learning*, *46*(1), 351–360.
Keerthi, S., & Lin, C. (2003). Asymptotic behaviors of support vector machines with Gaussian kernel. *Neural Computation*, *15*(7), 1667–1689.
Keerthi, S., Shevade, S., Bhattacharyya, C., & Murthy, K. (2001). Improvements to Platt's SMO algorithm for SVM classifier design. *Neural Computation*, *13*(3), 637–649.
Koltchinskii, V. (2001). Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, *47*(5), 1902–1914.
Koltchinskii, V. (2006). Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, *34*(6), 2593–2656.
Lawler, E., & Wood, D. (1966). Branch-and-bound methods: A survey. *Operations Research*, *14*, 699–719.
Lee, W., Bartlett, P., & Williamson, R. (1998). The importance of convexity in learning with squared loss. *IEEE Transactions on Information Theory*, *44*(5), 1974–1980.
Martein, L., & Schaible, S. (1987). On solving a linear program with one quadratic constraint. *Decisions in Economics and Finance*, *10*(1), 75–90.
McDiarmid, C. (1989). On the method of bounded differences. *Surveys in Combinatorics*, *141*(1), 148–188.
Mendelson, S. (2003). On the performance of kernel classes. *The Journal of Machine Learning Research*, *4*, 759–771.
Milenova, B., Yarmus, J., & Campos, M. (2005). SVM in oracle database 10g: Removing the barriers to widespread adoption of support vector machines. In *International conference on very large data bases* (pp. 1152–1163).
Morozov, V., Nashed, Z., & Aries, A. (1984). *Methods for solving incorrectly posed problems*. New York: Springer.
Munder, S., & Gavrila, D. (2006). An experimental study on pedestrian classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(11), 1863–1868.
Oneto, L., Ghio, A., Ridella, S., & Anguita, D. (2014). Fully empirical and data-dependent stability-based bounds. *IEEE Transactions on Cybernetics*. doi:10.1109/TCYB.2014.2361857.

Park, M. Y., & Hastie, T. (2007). L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *69*(4), 659–677.
Pelckmans, K., Suykens, J., & De Moor, B. (2004) Morozov, Ivanov and Tikhonov regularization based ls-SVMs. In *Neural information processing systems* (pp. 1216–1222).
Platt, J. (1998). Sequential minimal optimization: A fast algorithm for training support vector machines. *Advances in Kernel Methods—Support Vector Learning*, *208*, 1–21.
Platt, J. (1999). Using analytic QP and sparseness to speed training of support vector machines. In *Advances in neural information processing systems* (pp. 557–563).
Poggio, T., Mukherjee, S., Rifkin, R., Rakhlin, A., & Verri, A. (2002). b. In *Uncertainty in geometric computations*.
Poggio, T., Rifkin, R., Mukherjee, S., & Niyogi, P. (2004). General conditions for predictivity in learning theory. *Nature*, *428*(6981), 419–422.
Pontil, M., & Verri, A. (1998). Properties of support vector machines. *Neural Computation*, *10*(4), 955–974.
Schölkopf, B. (2001). The kernel trick for distances. In *Neural information processing systems* (pp. 301–307).
Schölkopf, B., Herbrich, R., & Smola, A. (2001). A generalized representer theorem. In *Computational learning theory* (pp. 416–426).
Shawe-Taylor, J., Bartlett, P. L., Williamson, R. C., & Anthony, M. (1998). Structural risk minimization over data-dependent hierarchies. *IEEE Transactions on Information Theory*, *44*(5), 1926–1940.
Shawe-Taylor, J., & Cristianini, N. (2004). *Kernel methods for pattern analysis*. Cambridge: Cambridge University Press.
Shawe-Taylor, J., & Sun, S. (2011). A review of optimization methodologies in support vector machines. *Neurocomputing*, *74*(17), 3609–3618.
Suykens, J., & Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural processing letters*, *9*(3), 293–300.
Tikhonov, A., Arsenin, V., & John, F. (1977). *Solutions of ill-posed problems*. Washington, DC: Winston.
Tomasi, C. (2004). Learning theory: Past performance and future results. *Nature*, *428*(6981), 378–378.
Vapnik, V. (1998). *Statistical learning theory*. New York: Wiley.
Vapnik, V. (2000). *The nature of statistical learning theory*. New York: Springer.
Vorontsov, K. (2010). Exact combinatorial bounds on the probability of overfitting for empirical risk minimization. *Pattern Recognition and Image Analysis*, *20*(3), 269–285.
Yuille, A., & Rangarajan, A. (2003). The concave–convex procedure. *Neural Computation*, *15*(4), 915–936.