Regularization Techniques and Suboptimal Solutions to Optimization Problems in Learning from Data

Giorgio Gnecco
giorgio.gnecco@dist.unige.it
Departments of Communications, Computer, and System Sciences and of Computer and Information Science, University of Genova, Genova 16146, Italy

Marcello Sanguineti
marcello@dist.unige.it
Department of Communications, Computer, and System Sciences, University of Genova, Genova 16146, Italy

Various regularization techniques are investigated in supervised learning from data. Theoretical features of the associated optimization problems are studied, and sparse suboptimal solutions are searched for. Rates of approximate optimization are estimated for sequences of suboptimal solutions formed by linear combinations of \( n \)-tuples of computational units, and statistical learning bounds are derived. As hypothesis sets, reproducing kernel Hilbert spaces and their subsets are considered.

1 Introduction

For nonempty sets \( X \) and \( Y, Y \subset \mathbb{R} \), a joint probability measure \( \rho \) on \( X \times Y \), and a loss function \( V : \mathbb{R}^2 \rightarrow [0, +\infty) \) such that \( V(u, u) = 0 \) for all \( u \in \mathbb{R} \), statistical learning theory (SLT) (Cucker & Smale, 2001; Vapnik, 1998) models the learning problem as the minimization of the expected error functional \( \mathcal{E}_V(f) = \int_{X \times Y} V(f(x), y) d\rho \) over a suitable linear space \( \mathcal{H} \) of functions, called hypothesis space. The quantity \( V(f(x), y) \) measures how much is lost when \( f(x) \) is computed instead of \( y \); a typical choice is the square loss \( V(f(x), y) = (f(x) - y)^2 \). The mapping \( f_\rho : X \rightarrow \mathbb{R} \), defined as \( f_\rho(x) = \int_{\mathbb{R}} y d\rho(y | x) \), is called regression function; if \( f_\rho \in \mathcal{H} \) and the square loss is used, then \( f_\rho \) is the unique minimizer over \( \mathcal{H} \) of the expected error functional (Cucker & Smale, 2001).

In practice, since \( \rho \) usually is unknown or difficult to model, the minimization of the expected error functional is replaced by the minimization of a suitable empirical approximation, which depends on a data sample \( z = \{(x_i, y_i) \in X \times Y, i = 1, \ldots, m\} \), also called the training set. Typically the data sample is made up of \( m \) independent and identically distributed (i.i.d.) pairs \((x_i, y_i)\), generated according to the probability measure \( \rho \). The empirical error functional is defined as \( \mathcal{E}_{z, V}(f) = \frac{1}{m} \sum_{i=1}^{m} V(f(x_i), y_i) \). To simplify
the notation, we denote by $E$ and $E_z$ the expected and empirical error functionals, respectively, with the square loss function, that is, we let

$$E(f) = \int_{X \times Y} (f(x) - y)^2 \, d\rho \quad \text{and} \quad E_z(f) = \frac{1}{m} \sum_{i=1}^{m} (f(x_i) - y_i)^2.$$ 

For suitable choices of subsets of the hypothesis space $\mathcal{H}$, various regularization techniques (Tikhonov & Arsenin, 1977), interpreted according to SLT (Evgeniou, Pontil, & Poggio, 2000), allow one to derive upper bounds on the difference between the expected and empirical errors. The bounds hold uniformly over such subsets of the hypothesis space, with high a priori probability with respect to the random draw of the data sample $z$; in the following, we shall refer to them as SLT bounds.

As an example of a regularization technique, one can add to the empirical error functional one term, called a stabilizer, which penalizes undesired properties of a candidate solution. For example, when $\mathcal{H}$ is a linear space endowed with the norm $\|\cdot\|_{\mathcal{H}}$, one may search for a minimizer of the regularized empirical error functional $E_z(f) + \gamma \|f\|_{\mathcal{H}}^2$, where $\gamma > 0$ is called the regularization parameter, thus considering the problem

$$\min_{f \in \mathcal{H}} (E_z(f) + \gamma \|f\|_{\mathcal{H}}^2). \quad (1.1)$$

The parameter $\gamma$ controls the trade-off between the two requirements of (1) fitting to the data sample (via the value $E_z(f)$ of the empirical error in correspondence of $f$) and (2) penalizing every candidate solution $f$ with a large value of the squared norm $\|f\|_{\mathcal{H}}^2$. For suitable choices of the hypothesis space $\mathcal{H}$, the norm $\|f\|_{\mathcal{H}}$ is a measure of the smoothness of the function (see, e.g., Girosi, 1998). Alternatively, one can restrict the minimization of the empirical error to a subset $M$ of $\mathcal{H}$, which is called the hypothesis set and contains only functions with a desired behavior, so one considers the problem

$$\min_{f \in M} E_z(f). \quad (1.2)$$

A third possibility is to combine the two methods (use of a stabilizer and restriction to a hypothesis set). These three approaches are applications to learning from data of regularization techniques for ill-posed inverse problems developed during the 1960s. The first is an application of Tikhonov regularization, the second and the third of Ivanov and Miller regularizations, respectively. A related form of regularization is Phillips regularization (Bertero, 1989).

---

1Whenever we write “min,” we implicitly suppose that such a minimum exists. If it does not, we mean that we are interested, for $\varepsilon > 0$, in an $\varepsilon$-near minimum point. For example, when the minimum in equation 1.1 is not achieved, we are interested in $f_\varepsilon \in \mathcal{H} : E_z(f_\varepsilon) + \gamma \|f_\varepsilon\|_{\mathcal{H}}^2 \leq \inf_{f \in \mathcal{H}} (E_z(f) + \gamma \|f\|_{\mathcal{H}}^2) + \varepsilon$. 

\[\inf_{f_\varepsilon \in \mathcal{H}} (E_z(f_\varepsilon) + \gamma \|f_\varepsilon\|_{\mathcal{H}}^2) + \varepsilon. \]
When the hypothesis space is a Hilbert space of a special type, that is, a reproducing kernel Hilbert space (Aronszajn, 1950), the optimal solutions to the learning problems associated with Tikhonov, Ivanov, Phillips, and Miller regularizations are expressed in terms of linear combinations of the columns of an $m \times m$ matrix obtained on the basis of the data sample (see theorem 1). Unfortunately, in general, the number of nonzero coefficients in such linear combinations is not much smaller than the size $m$ of the data sample; in other words, such optimal solutions are not sparse. A sparse suboptimal solution $f$ has several advantages over an optimal nonsparse one $f^o$; for example, storing its coefficients requires less memory and computing $f(x)$ for $x$ not belonging to the data sample requires less time than computing $f^o(x)$. Motivated by these advantages, we investigate rates of approximate optimization achievable by sparse suboptimal solutions for the regularization techniques already identified.

The letter is both a collection of original contributions and a survey of related results. Propositions 1(iv), 3, 4, and 6 and the results in sections 4, 5, and 6 are original contributions. In proposition 1(iv), we prove the existence of a solution to Miller regularization in a slightly more general setting than the one considered in Miller (1970). Propositions 3 and 4 provide lower and upper bounds on the parameters in the Ivanov-regularized learning problem. Proposition 6 states a simple but useful property of sparse suboptimal solutions to the four regularized learning problems. Section 4 improves rates of approximate optimization by sparse suboptimal solutions derived in Kůrková and Sanguineti (2005b) and Zhang (2002a) for Tikhonov regularization. Sections 5 and 6 give new estimates for Ivanov, Miller, and Phillips regularization in learning from data. For all these techniques, we also investigate SLT bounds for sparse suboptimal solutions: we provide conditions under which one can control with high probability (uniformly over the family of functions in which one searches for suboptimal solutions) the difference between the expected and the empirical errors.

As to the survey contribution of the letter, part of section 2 summarizes known relationships among Tikhonov, Ivanov, Phillips, and Miller regularizations. The first part of section 3 discusses some drawbacks to their respective optimal solutions that motivate the search for sparse suboptimal solutions. Section 7 provides a short discussion on some available algorithms, which can be used to find sparse suboptimal solutions to the regularized learning problems.

To make the letter self-contained, in the appendix, we summarize some tools from SLT, exploited in sections 4 to 6.

2 Comparison of Regularization Techniques

Reproducing kernel Hilbert spaces (RKHSs), defined almost sixty years ago (Aronszajn, 1950), are now commonly used as hypothesis spaces in learning theory. Combined with regularization techniques, often they allow
good generalization capabilities of the learned models and enforce desired
smoothness properties of the solutions to the learning problems.

Given a nonempty set \( X \), as RKHS is a Hilbert space \( \mathcal{H}_K(X) \) made up
of functions on \( X \), such that for every \( u \in X \), the evaluation functional \( \mathcal{F}_u \),
defined for every \( f \in \mathcal{H}_K(X) \) as \( \mathcal{F}_u(f) = f(u) \), is bounded (Aronszajn, 1950;
Berg, Christensen, & Ressel, 1984; Cucker & Smale, 2001). We consider real
RKHSs. By the Riesz representation theorem (Friedman, 1982), for every
\( u \in X \), there exists a unique element \( K_u \in \mathcal{H}_K(X) \) such that for every \( f \in \mathcal{H}_K(X) \), one has
\[
\mathcal{F}_u(f) = \langle f, K_u \rangle_K,
\]
where \( \langle \cdot, \cdot \rangle_K \) denotes the inner product in \( \mathcal{H}_K(X) \); this is called the reproducing property. The symmetric function \( K : X \times X \to \mathbb{R} \) defined for every \( u, v \in X \) as \( K(u, v) = \langle K_u, K_v \rangle_K \), is called the kernel associated with
\( \mathcal{H}_K(X) \). The kernel \( K \) is a positive-semidefinite (psd) function; for all positive integers \( m \), all \( (w_1, \ldots, w_m) \in \mathbb{R}^m \), and all \( (u_1, \ldots, u_m) \in X^m \), one has
\[
\sum_{i,j=1}^m w_i w_j K(u_i, u_j) \geq 0.
\]
The kernel is called positive definite (pd) if the equality in the previous definition holds only for \( (w_1, \ldots, w_m) = (0, \ldots, 0) \), when all the \( u_i \) are different. By equation 2.1, \( K(u, \cdot) \in \mathcal{H}_K(X) \), and it coincides with \( K_u \).

Typical regularization techniques perturb the functional (Tikhonov regularization); restrict the minimization to a subset of the hypothesis space, called the hypothesis set (Ivanov regularization); replace the empirical error with another functional and use the former to define the hypothesis set (Phillips regularization); or combine the last two possibilities (Miller regularization). (See Table 1.) As Tikhonov (Tikhonov & Arsenin, 1977) developed a unifying formulation of regularization theory, sometimes all such techniques are improperly gathered as “Tikhonov regularization.”

We use the following standard notations from optimization theory (see,
e.g., Dontchev & Zolezzi, 1993). Given a function space \( \mathcal{X} \), a set \( M \subseteq \mathcal{X} \),
and a functional \( \Phi : M \to \mathbb{R} \), we denote by \( (M, \Phi) \) the problem \( \inf_{f \in M} \Phi(f) \).
Every \( f^* \in M \) such that \( \Phi(f^*) = \min_{f \in M} \Phi(f) \) is called a solution or a minimum point of the problem \( (M, \Phi) \). We denote by
\[
\text{argmin}(M, \Phi) = \left\{ f \in M : \Phi(f) = \min_{f \in M} \Phi(f) \right\}
\]
the set of solutions of \( (M, \Phi) \). A sequence \( \{f_n\} \) of elements of \( M \) is called \( \Phi \)-minimizing over \( M \) if \( \lim_{n \to +\infty} \Phi(f_n) = \inf_{f \in M} \Phi(f) \).

Tikhonov regularization in a strict sense (see row 1 of Table 1) involves
the minimization of the functional \( \mathcal{E}_K(\cdot) + \gamma \| \cdot \|_K^2 \) over the whole RKHS
\( \mathcal{H}_K(X) \). Besides motivations based on SLT (Evgeniou et al., 2000), which we shall discuss later, this kind of regularization is motivated by the fact that the norms \( \| \cdot \|_K \) on suitable RKHSs play the role of measures of various
### Table 1: Regularization Techniques for Learning from Data in RKHSs.

<table>
<thead>
<tr>
<th>Regularization Technique</th>
<th>Regularization Parameter</th>
<th>Functional</th>
<th>Hypothesis Set</th>
<th>Minimization Problem</th>
<th>Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tikhonov</td>
<td>$\gamma &gt; 0$</td>
<td>$\Phi_{T,\gamma}(\cdot) = \mathcal{E}<em>\mathcal{Z}(\cdot) + \gamma | \cdot |^2</em>{\mathcal{K}}$</td>
<td>$\mathcal{H}_\mathcal{K}(X)$</td>
<td>Problem $T_{\gamma}$; $(\mathcal{H}<em>\mathcal{K}(X), \Phi</em>{T,\gamma})$</td>
<td>$f_{T,\gamma}^0$</td>
</tr>
<tr>
<td>Ivanov</td>
<td>$r &gt; 0$</td>
<td>$\Phi_{I,r}(\cdot) = \mathcal{E}_\mathcal{Z}(\cdot)$</td>
<td>$B_r(| \cdot |_{\mathcal{K}})$</td>
<td>Problem $I_r$; $(B_r(| \cdot |<em>{\mathcal{K}}), \Phi</em>{I,r})$</td>
<td>$f_{I,r}^0$</td>
</tr>
<tr>
<td>Phillips</td>
<td>$\eta \geq 0$</td>
<td>$\Phi_{P,\eta}(\cdot) = | \cdot |^2_{\mathcal{K}}$</td>
<td>$F_{\mathcal{Z},\eta} = { f \in \mathcal{H}<em>\mathcal{K}(X) : \mathcal{E}</em>\mathcal{Z}(f) \leq \eta^2 }$</td>
<td>Problem $P_{\eta}$; $(F_{\mathcal{Z},\eta}, \Phi_{P,\eta})$</td>
<td>$f_{P,\eta}^0$</td>
</tr>
<tr>
<td>Miller</td>
<td>$r &gt; 0$, $\eta \geq 0$</td>
<td>$\Phi_{M,r,\eta}(\cdot) = \mathcal{E}<em>\mathcal{Z}(\cdot) + \left( \frac{\eta}{r} \right)^2 | \cdot |^2</em>{\mathcal{K}}$</td>
<td>$G_{\mathcal{Z},r,\eta}$</td>
<td>Problem $M_{r,\eta}$; $(G_{\mathcal{Z},r,\eta}, \Phi_{M,r,\eta})$</td>
<td>$f_{M,r,\eta}^0$</td>
</tr>
</tbody>
</table>
types of oscillations of input-output mappings (Girosi, 1998). Ivanov regularization (row 2 of Table 1) considers the minimization of the empirical error functional \( E_z \) on a closed ball \( B_r(\|\cdot\|_K) \) of radius \( r \) in the RKHS \( \mathcal{H}_K(X) \). In Phillips regularization (row 3), the functionals \( E_z \) and \( \|\cdot\|_K^2 \) exchange their roles: the minimization of \( E_z \) is replaced by the minimization of \( \|\cdot\|_K^2 \), and, in turn, an upper bound on the values of \( E_z \) defines the hypothesis set. Finally, Miller regularization (row 4) combines the approaches of Ivanov and Phillips regularizations: both the empirical error functional and the squared-norm functional contribute to defining the regularized functional and the hypothesis set.

The four techniques consider regularization of the learning problem from different points of view; each has pros and cons from an optimization perspective. One advantage of Tikhonov regularization is that it merely requires unconstrained optimization algorithms, whereas the other three methods need algorithms for constrained optimization. However, the choice of the regularization parameter \( \gamma \) is not straightforward, even though various criteria have been proposed for such a choice (see, e.g., Bertero, 1989; Wahba, 1990; and Engl, Hanke, & Neubauer, 2000, for general inverse problems and Cucker & Smale, 2002, and De Vito, Rosasco, Caponnetto, De Giovannini, & Odone, 2005, more specifically for learning problems). From a practical point of view, the regularization parameters \( r > 0 \) and \( \eta \geq 0 \) used in Ivanov, Phillips, and Miller regularization methods have a more direct interpretation: \( r \) and \( \eta \) play the role of upper bounds on the smoothness of the solution and the mean square error on the data, respectively. Their choices may also be guided by the equivalence of the four regularization techniques, stated in proposition 2. (We do not consider the case \( r = 0 \), as the corresponding hypothesis set consists of only the null function \( f = 0 \). The case \( \eta = 0 \) corresponds to a hypothesis set that interpolates the data exactly.)

In the rest of this section, we summarize some results on existence, uniqueness, and properties of the solutions to the regularized problems. Then we shall give some original contributions (see propositions 1(iv), 3, and 4).

Proposition 1 investigates the existence and uniqueness of the solutions to problems \( T_y, I_r, P_\eta, \) and \( M_{r,\eta} \). Using standard notations (Engl et al., 2000; Kúrková, 2005), we denote by \( f_z^\dagger \) the minimum-norm solution to the problem \((\mathcal{H}_K(X), E_z)\) (also called normal pseudosolution) and by \( L_x \) the operator \( L_x : \mathcal{H}_K(X) \to \mathbb{R}^m \) defined for every \( f \in \mathcal{H}_K(X) \) as

\[
L_x(f) = \left( \frac{f(x_1)}{\sqrt{m}}, \ldots, \frac{f(x_m)}{\sqrt{m}} \right)^T,
\]

(2.2)

where the superscript \(^T\) denotes transposition. For \( y = (y_1, \ldots, y_m)^T \in \mathbb{R}^m \), \( Py \) is the component of \( y \) in the range of the operator \( L_x \) and \( Qy \), its
orthogonal component. As in Table 1, we denote by $G_{z,r,\eta} = B_r(\|\cdot\|_K) \cap \{f \in \mathcal{H}_K(X) : \mathcal{E}_z(f) \leq \eta^2\}$ the hypothesis set of Miller regularization.

**Proposition 1 (existence and uniqueness of the solutions).** Let $X$ be a nonempty set, $K : X \times X \to \mathbb{R}$ a psd kernel, $m$ a positive integer, and $z$ a data sample of size $m$. For every $\gamma, r > 0$ and $\eta \geq 0$, the following hold:

i. Problem $T_\gamma$ has a unique solution $f^o_{T,\gamma}$.

ii. Problem $I_r$ has a solution $f^o_{I,r}$, which is unique if $r \leq \|f^*_z\|_K$.

iii. Problem $P_\eta$ has a solution $f^o_{P,\eta}$, which is unique if $\|Qy\|_2^2/m \leq \eta^2$. For $\|Qy\|_2^2/m \leq \eta^2$ one has $f^o_{P,\eta} = 0$.

iv. Problem $M_{r,\eta}$ has a solution $f^o_{M,\eta}$ if the set $G_{z,r,\eta}$ is nonempty; the solution is unique if the set $G_{z,\sqrt{r},\sqrt{\eta}}$ is nonempty too.

**Proof.** A proof of proposition 1(i) together with an explicit expression for $f^o_{T,\gamma}$, is given in Engl et al. (2000).

ii, iii. These follow by (Bertero, 1989).

iv. As our formulation of Miller regularization slightly differs from the original one in Miller (1970), we provide the proof of the existence of $f^o_{M,\eta}$ when the set $G_{z,r,\eta}$ is nonempty. For the proof of existence and uniqueness in the case $G_{z,\sqrt{r},\sqrt{\eta}} \subset G_{z,r,\eta}$, one can apply the arguments of Miller (1970).

Let $\{f_n\}$ be a minimizing sequence for the problem $(G_{z,r,\eta}, \Phi_{M,\eta})$. By definition, for every positive integer $n$, one has $\mathcal{E}_z(f_n) \leq \eta^2$ and $f_n \in B_r(\|\cdot\|_K)$, so $\{f_n\}$ is a bounded sequence. Thus, there exist a subsequence $\{f_{n_h}\}$ and an element $f^* \in \mathcal{H}_K(X)$ such that $\{f_{n_h}\} \to f^*$ as $h \to +\infty$, where “$\to$” denotes weak convergence. As by definition in an RKHS, all evaluation functionals are bounded; this implies $L_x(f_{n_h}) \to L_x(f^*)$ as $h \to +\infty$, so $\mathcal{E}_x(f) \leq \eta^2$. By the weak lower semicontinuity of the norm in a Hilbert space (Rauch, 1991), it follows that $\|f^*\|_K \leq \inf_{h \in \mathbb{N}} \|f_{n_h}\|_K$, so $f^* \in B_r(\|\cdot\|_K)$. Then $f^* \in G_{z,r,\eta}$, and it is a solution to problem $M_{r,\eta}$.

For ease of reference, we collect in the following assumption the conditions that guarantee the existence and uniqueness of the solutions to the four regularized problems. We further assume $\|f^*_z\|_K > 0$, as it is usually the case (otherwise, $y_i = 0$ for $i = 1, \ldots, m$):

**Assumption 1.** Let the following hold:

- Problem $T_\gamma$: $\gamma > 0$.
- Problem $I_r$: $0 < r \leq \|f^*_z\|_K$.
- Problem $P_\eta$: $0 \leq \|Qy\|_2^2/m \leq \eta^2$.
- Problem $M_{r,\eta}$: $r > 0$ and $\eta \geq 0$ such that $G_{z,\sqrt{r},\sqrt{\eta}}$ is nonempty.
Under assumption 1, Tikhonov, Ivanov, Phillips, and Miller regularizations are equivalent in the sense specified by proposition 2:

**Proposition 2 (equivalent regularizations).** Let $X$ be a nonempty set, $K : X \times X \to \mathbb{R}$ a psd kernel, $m$ a positive integer, and $z$ a data sample of size $m$. Then under assumption 1, the solutions to problems $T_\gamma, I_r, P_\eta,$ and $M_\eta$ are equivalent in the following sense:

i. $f^o_{T,\gamma} \in \arg\min (H_K(X), \Phi_{T,\gamma}) \implies f^o_{T,\gamma} \in \arg\min (B_r(\gamma)(\| \cdot \|_K), \Phi_{I,r}(\gamma))$,
where $r(\gamma) = \| f^o_{T,\gamma} \|_K$.

ii. $f^o_{T,\gamma} \in \arg\min (H_K(X), \Phi_{T,\gamma}) \implies f^o_{T,\gamma} \in \arg\min (F_z(\gamma)(\| \cdot \|_K), \Phi_{P,\eta}(\gamma))$,
where $\eta(\gamma) = \sqrt{\mathcal{E}_z(f^o_{T,\gamma})}$.

iii. $f^o_{T,\gamma} \in \arg\min (H_K(X), \Phi_{T,\gamma}) \implies f^o_{T,\gamma} \in \arg\min (G_{z,\eta,r}(\gamma), \Phi_{M,\eta,r}(\gamma))$,
where $\gamma = \left(\frac{\eta}{r}\right)^2$, $\eta \geq \sqrt{\frac{2\mathcal{E}_z(f^o_{T,\gamma})}{m}}$, and $r \geq \sqrt{2\| f^o_{T,\gamma} \|_K}$.

For $\gamma \in (0, +\infty)$, the functions $r(\gamma)$ and $\eta(\gamma)$ in items i and ii are one-to-one and monotonic, with ranges $(0, \| f^o_{T,\gamma} \|_K)$ and $(\| f^o_{T,\gamma} \|_K, \frac{\| f^o_{T,\gamma} \|_K}{\sqrt{m}})$, respectively.

**Proof.** See Bertero (1989), Vasin (1970), or Mukherjee, Rifkin, & Poggio (2002) (the last two references consider only cases i and ii). The last statement, about the properties of $r(\gamma)$ and $\eta(\gamma)$, follows the expressions of $\| f^o_{T,\gamma} \|_K$ and $\mathcal{E}_z(f^o_{T,\gamma})$ given in Bertero (1989).

For the case of Tikhonov regularization, the next theorem is known as the representer theorem. It was originally proven in Wahba (1990), later in Cucker and Smale (2002) using Fréchet derivatives, and then in Kůrková (2004, 2005) via properties of inverse problems and the operator $L_x$ (see equation 2.2).

We denote by $K[x]$ the Gram matrix of the data with respect to $x$—the $m \times m$ matrix whose elements are $K[x]_{i,j} = K(x_i, x_j)$.

**Theorem 1 (representer theorem).** Let $X$ be a nonempty set, $K : X \times X \to \mathbb{R}$ a psd kernel, $m$ a positive integer, and $z$ a data sample of size $m$. Then under assumption 1, the optimal solutions to problems $T_\gamma, I_r, P_\eta,$ and $M_\eta$ are of the form

$$f^o_z(\cdot) = \sum_{i=1}^{m} c_{\alpha,i} K_{x_i}(\cdot), \quad (2.3)$$

where, for every $m$, every $K[x]$, and every $y$, the values of $\alpha > 0$ corresponding to $f^o_{T,\gamma}$, $f^o_{I,r}$, $f^o_{P,\eta}$, and $f^o_{M,\eta}$ are functions of the respective parameters $\gamma$, $r$, $\eta$, and $(r, \eta)$. The vector $c_\alpha = (c_{\alpha,1}, \ldots, c_{\alpha,m})^T$ is the unique solution to the well-posed linear system

$$\left(\alpha m I + K[x]\right)c_\alpha = y. \quad (2.4)$$
For problem $T_\gamma$ one has $\alpha(\gamma) = \gamma$ and for problem $P_\eta$, the case $\frac{\|y\|_2^2}{m} \leq \eta^2$ (see proposition 1(iii) is covered by letting $\alpha \to +\infty$.

**Proof.** By the properties of continuous linear operators (Bertero, 1989; Groetch, 1977), for every $u \in \mathbb{R}^m$, the adjoint operator $L^* : \mathbb{R}^m \to \mathcal{H}_K(X)$ is given by $L^*(u) = \frac{1}{\sqrt{m}} \sum_{i=1}^m u_i K_{x_i}$ (see also Kůrková, 2005). So for the case of Tikhonov regularization, equation 2.5 follows Bertero (1989), adapting to $L^*$ the expression for $f^\alpha_z$ given in equation 226 in Bertero. The other cases follow by the equivalence of the four regularizations stated in proposition 2.

Despite the equivalence, in the sense stated by proposition 2, among Tikhonov, Ivanov, Phillips, and Miller regularizations, the values of the parameters guaranteeing such equivalence are not known a priori. For example, the function $r(\gamma)$ in proposition 2 depends on the solution $f_{oT,\gamma}$.

In the remainder of the letter, in general we shall not make the assumption that $r(\gamma), \eta(\gamma)$ and their inverses, $\gamma(r)$ and $\gamma(\eta)$, are known, and we shall exploit lower and upper bounds on their values. The following proposition provides such bounds on $\gamma(r)$:

**Proposition 3.** Let $X$ be a nonempty set, $K : X \times X \to \mathbb{R}$ a psd kernel, $m$ a positive integer, $z$ a sample of data of size $m$, $0 < r \leq \|f^\dagger_z\|_K$, and $\lambda_{\min}, \lambda_{\max}$ the minimum and maximum eigenvalues of $K[x]$, respectively. Then the following hold:

i. $\frac{\sqrt{\lambda_{\min}} \|y\|_2 - \lambda_{\max} r}{mr} \leq \gamma(r) \leq \frac{\sqrt{\lambda_{\max}} \|y\|_2 - \lambda_{\min} r}{mr}$.

ii. $\min_{\lambda \in [\lambda_{\min}, \lambda_{\max}]} \sqrt{\lambda} \frac{\|y\|_2 - \lambda r}{mr} \leq \gamma(r) \leq \max_{\lambda \in [\lambda_{\min}, \lambda_{\max}]} \sqrt{\lambda} \frac{\|y\|_2 - \lambda r}{mr}$.

iii. Let $\lambda^* = \frac{\|y\|_2}{4m^2 r}$. If $\lambda^* \in [\lambda_{\min}, \lambda_{\max}]$, then $\min_{\lambda \in [\lambda_{\min}, \lambda_{\max}]} \sqrt{\lambda} \frac{\|y\|_2 - \lambda r}{mr} \leq \gamma(r) \leq \max_{\lambda \in [\lambda_{\min}, \lambda_{\max}]} \sqrt{\lambda} \frac{\|y\|_2 - \lambda r}{mr}$.

**Proof.** (i) For $0 < r \leq \|f^\dagger_z\|_K$, by proposition 2(i), one has $f_{T,\gamma(r)}^0 = f_{I,r}$ and $\|f_{T,\gamma(r)}^0\|_K = r$. By theorem 1, we get $f_{T,\gamma(r)}^0 = \sum_{i=1}^m c_i K_{x_i}$, where

$$c = (c_1, \ldots, c_m)^T = (\gamma(r) m I + K[x])^{-1} y,$$

and by the reproducing property 2.1, $\|f_{T,\gamma(r)}^0\|_K = \sqrt{c^T K[x] c}$. Then the statement follows by

$$r = \|f_{T,\gamma(r)}^0\|_K = \sqrt{c^T K[x] c} \leq \sqrt{\lambda_{\max}} \|c\|_2 \leq \frac{\sqrt{\lambda_{\max}}}{\gamma(r) m + \lambda_{\min}} \|y\|_2. \quad (2.5)$$
and, similarly,
\[ r \geq \sqrt{\lambda_{\text{min}}} \|\mathbf{c}\|_2 \geq \frac{\sqrt{\lambda_{\text{min}}}}{\gamma(r)m + \lambda_{\text{max}}} \|\mathbf{y}\|_2. \] (2.6)

(ii) Proceeding as in the proof of case i, since the matrices \(K[\mathbf{x}]\) and \((\gamma(r)m I + K[\mathbf{x}])^{-1}\) have the same eigenvectors, a slightly refined analysis gives
\[ r = \sqrt{\mathbf{c}^T K[\mathbf{x}] \mathbf{c}} = \sqrt{\mathbf{y}^T (\gamma(r)m I + K[\mathbf{x}])^{-T} K[\mathbf{x}] (\gamma(r)m I + K[\mathbf{x}])^{-1} \mathbf{y}} \]
\[ \leq \max_{\lambda \in [\lambda_{\text{min}}, \lambda_{\text{max}}]} \frac{\sqrt{\lambda}}{\gamma(r) m + \lambda} \|\mathbf{y}\|_2 \] (2.7)
and
\[ r \geq \min_{\lambda \in [\lambda_{\text{min}}, \lambda_{\text{max}}]} \frac{\sqrt{\lambda}}{\gamma(r)m + \lambda} \|\mathbf{y}\|_2. \] (2.8)

We conclude by rewriting the last two inequalities in terms of \(\gamma(r)\).

(iii) The estimates follow from case ii by showing that the function \(f : [0, +\infty) \to \mathbb{R}\) defined as \(f(\lambda) = \sqrt{\lambda} \|\mathbf{y}\|_2 - \lambda r\) is strictly increasing in \([0, \lambda^*]\), strictly decreasing in \([\lambda^*, +\infty)\), and such that \(f(0) = 0\).

Note that the first bound in proposition 3(i) is meaningful only when \(0 < r < \frac{\sqrt{\lambda_{\text{min}}}}{\lambda_{\text{max}}} \|\mathbf{y}\|_2\). The first bound in the proposition is slightly better, but still not meaningful if \(\lambda_{\text{min}} = 0\). However, there exist cases in which \(\lambda_{\text{min}}\) can be bounded from below away from 0. For example, for the gaussian kernel (defined for every \(\beta > 0\) and \((u, v) \in \mathbb{R}^d \times \mathbb{R}^d\) as \(K(u, v) = e^{-\beta \|u - v\|_2^2}\)), if the data samples have sufficiently separated input values \(x_1, \ldots, x_m\), then by a straightforward application of the Gershgorin circle theorem (Serre, 2002), all the eigenvalues of the Gram matrix \(K[\mathbf{x}]\) are concentrated around 1. For more general kernels, other lower bounds on \(\lambda_{\text{min}}\) (whose proofs are based on the Gershgorin circle theorem too) are given in Honeine, Richard, and Bermudez (2007).

When \(\lambda_{\text{min}} = 0\), the following proposition improves proposition 3. We denote by \(\lambda_{\text{min}}^+\) the minimum positive eigenvalue of the Gram matrix \(K[\mathbf{x}]\):

**Proposition 4.** Under the same assumptions of proposition 3, the same estimates given there hold if every instance of \(\lambda_{\text{min}}\) and \(\mathbf{y}\) is replaced by \(\lambda_{\text{min}}^+\) and \(\mathcal{P}\mathbf{y}\), respectively.

**Proof.** When \(\lambda_{\text{min}} > 0\), proposition 3 applies with no changes, as in this case \(\lambda_{\text{min}}^+ = \lambda_{\text{min}}\) and \(\mathcal{P}\mathbf{y} = \mathbf{y}\). When \(\lambda_{\text{min}} = 0\), let us decompose \(\mathbf{c}\) and \(\mathbf{y}\) as
\( c = c^N + c^R \) and \( y = y^N + y^R \), respectively, where \( c^N \) and \( y^N \) are the components of \( c \) and \( y \) in the null space of \( K[x] \) and \( c^R \) and \( y^R \) are the respective orthogonal components in the range of \( K[x] \). As \( K[x] \) and \( (\gamma(r) m I + K[x])^{-1} \) have the same eigenvectors, we get \( c^N = (\gamma(r) m I + K[x])^{-1} y^N \) and \( c^R = (\gamma(r) m I + K[x])^{-1} y^R \). By basic results of spectral theory, equation 2.5 becomes

\[
\begin{align*}
\gamma(r) &= \| f^o_{T, \gamma(r)} \|_K = \sqrt{c^T K[x] c} = \sqrt{(c^N)^T K[x] c^N + (c^R)^T K[x] c^R} \\
&= \sqrt{(c^R)^T K[x] c^R} = \sqrt{\lambda_{\max}} \| c^R \|_2 \leq \frac{\sqrt{\lambda_{\max}}}{\gamma(r) m + \lambda_{\min}} \| y \|_2.
\end{align*}
\]

One can proceed similarly for the other bounds. We conclude the proof noting that by proposition 3.1(iv) in Kúrková (2004), the operator \( L_x L_x^* : \mathbb{R}^m \to \mathbb{R}^m \) is represented in the canonical basis of \( \mathbb{R}^m \) by the matrix \( \frac{1}{m} K[x] \), from which it follows that \( \mathcal{P} y = y^R \).

The next proposition illustrates the relationships among the \( \| \cdot \|_K \)-norms of the solutions to problems \( I_r, P_\eta, \) and \( M_{r, \eta}, \) the corresponding minimum values of the empirical error functional, and the values of the parameters \( \gamma, r, \) and \( \eta \).

**Proposition 5.** Let \( X \) be a nonempty set, \( K : X \times X \to \mathbb{R} \) a psd kernel, \( m \) a positive integer, and \( z \) a data sample of size \( m \). Then under assumption 1, the following hold:

\[ \begin{align*}
\text{i. } & \| f^o_{P, \eta} \|_K \leq \| f^o_{M, \eta, r} \|_K \leq \| f^o_{I, r} \|_K \\
\text{ii. } & \mathcal{E}_z(f^o_{I, r}) \leq \mathcal{E}_z(f^o_{P, \eta}) \leq \mathcal{E}_z(f^o_{M, \eta, r}) \\
\text{iii. } & \gamma(r) \leq \left( \frac{\eta}{r} \right)^2 \leq \gamma(\eta),
\end{align*} \]

where \( \gamma(r) \) and \( \gamma(\eta) \) are the inverse of the functions \( r(\gamma) \) and \( \eta(\gamma) \), respectively, given in proposition 2.

**Proof.** Items i to iii, which were proven in Bertero (1989), follow by proposition 2.

Typically the norm \( \| \cdot \|_K \) in an RKHS is a measure of the smoothness of its functions (Girosi, 1998). So proposition 5 has the following interpretation. According to item i, for fixed values of the parameters \( \gamma, r, \) and \( \eta, \) the solution to the learning problem obtained by Phillips regularization is smoother than the solution obtained by Miller regularization, which in turn is smoother than the solution to the learning problem derived by Ivanov regularization. However, items i and ii imply a trade-off between smoothness and value of the empirical error functional (i.e., fitting to empirical data); roughly speaking, the smoother the solution, the larger the empirical error.
So the solution to problem $M_{r,\eta}$ has a degree of smoothness and a value of the empirical error that are intermediate between the corresponding values for the solutions to problems $P_{\eta}$ and $I_r$.

### 3 Optimal Versus Suboptimal Solutions

For a linear space $\mathcal{X}$ and $G \subset \mathcal{X}$, we denote by $\text{span}_n G$ the set of linear combinations of all $n$-tuples of elements of $G$:

$$\text{span}_n G = \left\{ \sum_{i=1}^{n} w_i g_i : w_i \in \mathbb{R}, \ g_i \in G \right\}.$$

By theorem 1, the solution $f^{\circ}_{T,\gamma}$ to problem $T_{\gamma}$ over the whole RKHS $\mathcal{H}_K(X)$ is a function in $\text{span}_{m} G_{K_x}$, where

$$G_{K_x} = \{ K(x_i, \cdot) : i = 1, \ldots, m \}.$$

As the cardinality of $G_{K_x}$ is equal to $m$, we have $\text{span}_{m} G_{K_x} = \text{span}_{m} G_{K_x}$, which can be interpreted as the set of all input-output functions of a computational model with one hidden layer of $m$ computational units computing functions in $G_{K_x}$. By proposition 2, under assumption 1, the solutions to the other three regularized methods can be obtained from instances of problem $T_{\gamma}$ with suitable choices of the regularization parameter $\gamma$. So the solutions to problems $I_r, P_{\eta},$ and $M_{r,\eta}$ are functions in $\text{span}_{m} G_{K_x}$ too. In particular, for the gaussian kernel, the solution given by theorem 1 can be computed by a gaussian radial basis function network (Girosi, 1994) with $m$ hidden units. However, for large values of $m$, such a network might not be practically implementable, and even when it is, one may wish to use a suboptimal but simpler network structure.

Although finding the optimal linear combination in $\text{span}_{m} G_{K_x}$ merely requires solving the linear systems of equations 2.4, practical applications are limited by the rates of convergence of iterative methods solving such linear systems. These rates depend on the size of the condition number of the corresponding matrices (see the discussion in Kůrková & Sanguineti, 2005b). Recall that the condition number of a nonsingular $m \times m$ matrix $A$ with respect to a norm $\| \cdot \|$ on $\mathbb{R}^m$ is defined as

$$\text{cond}(A) = \| A \| \| A^{-1} \|,$$

where $\| A \|$ denotes the norm of $A$ as a linear operator on $\{ \mathbb{R}^m, \| \cdot \| \}$. For every symmetric nonsingular $m \times m$ matrix $A$, $\text{cond}_2(A) = \frac{\lambda_{\text{max}}(A)}{\lambda_{\text{min}}(A)}$, where $\text{cond}_2(A)$ denotes the condition number of $A$ with respect to the $\| \cdot \|_2$-norm on $\mathbb{R}^m$ (Ortega, 1990).
In the discussion of this paragraph, to fix ideas, we consider problem \( T_\gamma \) and so in equation 2.4, we let \( \alpha = \gamma \). However, by proposition 2, similar remarks hold for the other three regularization techniques. By simple algebraic manipulations and spectral theory arguments, for the condition numbers of the matrices involved in solving the linear system 2.4, we get the following estimates (we assume for simplicity that \( \mathcal{K}[x] \) is positive definite, so that \( \lambda_{\text{min}} \neq 0 \):

\[
\text{cond}_2(\gamma m I + \mathcal{K}[x]) = \frac{\gamma m + \lambda_{\text{max}}}{\gamma m + \lambda_{\text{min}}} \leq \frac{\lambda_{\text{max}}}{\lambda_{\text{min}}} = \text{cond}_2(\mathcal{K}[x]). \tag{3.1}
\]

\[
\text{cond}_2(\gamma m I + \mathcal{K}[x]) = \frac{\gamma m}{\gamma m + \lambda_{\text{min}}} + \frac{\lambda_{\text{max}}}{\gamma m + \lambda_{\text{min}}} \leq 1 + \frac{\lambda_{\text{max}}}{\gamma m}. \tag{3.2}
\]

When \( \text{cond}_2(\mathcal{K}[x]) \) is sufficiently small, for every \( \gamma > 0 \), inequality 3.1 guarantees good conditioning of the matrix. However, for large values of the size \( m \) of the data sample, the matrix \( \mathcal{K}[x] \) may be ill conditioned. On the other hand, since \( \lim_{\gamma \to +\infty} (1 + \lambda_{\text{max}} / \gamma m) = 1 \), inequality 3.2 guarantees that the regularization parameter \( \gamma \) can be chosen such that \( \text{cond}_2(\gamma m I + \mathcal{K}[x]) \) is arbitrarily close to \( 1 \). Unfortunately, good conditioning of the matrices is not the only requirement for \( \gamma \), as its value must also allow a good fit to the empirical data and thus cannot be too large (an optimal choice of the regularization parameter \( \gamma \) is a compromise between these two opposing requirements).

In the following, we consider suboptimal solutions to problems \( T_\gamma, I_r, P_\eta, \) and \( M_{r,\eta} \) formed by linear combinations of \( n < m \) functions \( K_{x_1}, \ldots, K_{x_m} \) centered at the data points \( x_1, \ldots, x_m \), that is, by elements of \( \text{span}_n G_K \). So these suboptimal solutions have the structure of widespread connectionistic models, such as one-hidden-layer feedforward neural networks, and are sparse when \( n \ll m \). More generally, one can consider suboptimal solutions in \( \text{span}_n G_K \), where \( G_K = \{ K_x : x \in X \} \) (in this case, the kernel functions \( K_x \) are not necessarily centered at the input data points). The following are some motivations to search for a sparse suboptimal solution \( f^s \) instead of an optimal (usually nonsparse) one \( f^o \):

- Storing its coefficients requires less memory.
- The computational effort required to find \( f^s \) may be much smaller than for \( f^o \) (this happens, e.g., for greedy algorithms; see the discussion in section 7 and Zhang, 2002a, 2002b, 2003, for Tikhonov regularization).
- Once \( f^s \) has been found, computing \( f^s(x) \) for \( x \) not belonging to the data sample is easier than computing \( f^o(x) \).
- A sparse model has a better interpretability than a nonsparse one (it is a simpler model, which depends on fewer parameters; Zou & Hastie, 2005).
The next proposition states a simple but interesting property of approximate solutions from span $G_{K_x}$.

**Proposition 6.** Let $K$ be a psd kernel and $f \in \text{span } G_{K_x}$. Then $L_x(f) = 0$ if and only if $f = 0$.

**Proof.** As $f \in \text{span } G_{K_x}$, there exist $c_1, \ldots, c_m \in \mathbb{R}$ such that $f = \sum_{i=1}^{m} c_i K_{x_i}$. By the reproducing property 2.1, $f(x_i) = (K[x]c)_i$ and $\|f\|_K^2 = c^T K[x] c$. So for $f \in \text{span } G_{K_x}$, $f(x_i) = 0$ for all $i \in \{1, \ldots, m\}$ is equivalent to $\|f\|_K^2 = 0$. Proposition 6 implies that suboptimal solutions in span $n G_{K_x}$ have no components in the null space of the operator $L_x$; the same occurs (under the conditions of existence and uniqueness given by assumption 1) for the corresponding optimal solutions and the normal pseudosolution $f^\dagger_z$. The null space of $L_x$ represents a component of the data generation model that cannot be observed directly from the data (this motivates the terminology invisible component; (Bertero, 1989), so it cannot be recovered from the data without additional a priori information. Note also that by the reproducing property 2.1, the equality $(c^N)^T K[x] c^N = 0$ in the proof of proposition 4 implies $\sum_{i=1}^{m} c_i^N K_{x_i} = 0$. This is in accordance with proposition 6.

In sections 4 and 5, we estimate for Tikhonov and Ivanov regularizations the rates of approximate optimization achievable by sparse suboptimal solutions in span $n G_{K_x}$ with $n < m$ (without loss of generality, we let $n < m$, as for $n \geq m$ the optimal solutions belong to span $n G_{K_x}$). Phillips and Miller regularizations will be considered in section 6, where we also briefly address possible improvements allowed by the use of free internal parameters, that is, when suboptimal solutions are from span $n G_{K_x}, n < m$. For gaussian kernel units, the first situation corresponds to gaussians centered at $n < m$ data points, chosen among the sample input data $\{x_1, \ldots, x_m\}$. The second case corresponds to gaussians centered at $n < m$ points, to be chosen in the entire input set $X$.

The estimates will be expressed in terms of a norm called $G$-variation, denoted by $\| \cdot \|_G$ and defined for a subset $G$ of a normed linear space $(X, \| \cdot \|_X)$ as the Minkowski functional of the set $\text{cl conv } (G \cup -G)$ (Kúrková, 1997) (the closure is with respect to the norm $\| \cdot \|_X$ of $X$), that is, $\|f\|_G = \inf\{c > 0 : f/c \in \text{cl conv } (G \cup -G)\}$. $G$-variation is a generalization of the concept of total variation of integration theory (Kolmogorov & Fomin, 1975). (For properties of $G$-variation, see Kúrková, 1997; Kúrková & Sanguineti, 2001, 2002.)

We also use the concept of modulus of continuity. Let $S \subset X$ and $\Phi : S \to \mathbb{R}$ a continuous functional. For $f \in S$, the function $\alpha_f : [0, +\infty) \to [0, +\infty)$, defined as

$$\alpha_f(t) = \sup\{||\Phi(f) - \Phi(g)|| : g \in S, \|f - g\|_X \leq t\}.$$
is called the modulus of continuity of $\Phi$ at $f$. By definition, $\alpha_f$ is a nondecreasing function. By $\alpha_{T,y}$ and $\alpha_{I,r}$ we denote the moduli of continuity of $\Phi_{T,y}$ and $\Phi_{I,r}$, respectively, computed at the corresponding minimizers.

For $\varepsilon > 0$, $f_\varepsilon \in M$ is an $\varepsilon$-near minimum point of the problem $(M, \Phi)$ if $\Phi(f) \leq \inf_{f \in M} \Phi(f) + \varepsilon$.

4 Estimates for Tikhonov-Regularized Learning

For Tikhonov regularization, the following theorem improves the estimates of the accuracy of suboptimal solutions derived in Kůrková and Sanguineti (2005b). Using the same notation as in Makovoz (1996), we let

$$\xi_n(G_{K_x}) = \inf\{\xi > 0 : G_{K_x} \text{ can be covered by at most } n \text{ sets of diameter } \leq \xi\}.$$

**Theorem 2.** Let $X$ be a nonempty set, $K : X \times X \to \mathbb{R}$ a psd kernel with $s_K = \sup_{x \in X} \sqrt{K(x,x)} < +\infty$, $m$ a positive integer, $z$ a data sample of size $m$, and $\Delta_{T,y,n,z} = 2(\xi_{[m/2]}(G_{K_x})\|f^o_{T,y}\|_{G_{K_x}})^2$. Then for every positive integer $n < m$, the following hold.

i. Let $\alpha_{T,y}(t)$ be the modulus of continuity of $\Phi_{T,y}$ at $f^o_{T,y}$ and

$$d_{T,y,n,z} = \inf_{f \in \text{span}_n G_{K_x}} \|f^o_{T,y} - f\|_K.$$

Then

$$\inf_{f \in \text{span}_n G_{K_x}} \Phi_{T,y}(f) - \Phi_{T,y}(f^o_{T,y}) \leq \alpha_{T,y}(d_{T,y,n,z}), \quad (4.1)$$

where $\alpha_{T,y}(t) \leq (s_K^2 + \gamma)t^2$ and $d_{T,y,n,z} \leq \frac{\Delta_{T,y,n,z}}{n}$.

ii. Let $\varepsilon_n > 0$ and $f_n$ an $\varepsilon_n$-near minimum point of $(\text{span}_n G_{K_x}, \Phi_{T,y})$. Then

$$\|f_n - f^o_{T,y}\|_K^2 \leq \frac{(s_K^2 + \gamma)\Delta_{T,y,n,z} + \varepsilon_n}{\gamma n}. \quad (4.2)$$

**Proof.** (i) The inequality 4.1 follows by the definition of modulus of continuity. To prove the upper bound on $\alpha_{T,y}$, we proceed as follows. The first-order term in the Taylor’s expansion, centered at $f^o_{T,y}$, of the quadratic functional $\Phi_{T,y}$ vanishes by the first-order optimality condition for unconstrained optimization. So $\Phi_{T,y}(f) = \Phi_{T,y}(f^o_{T,y}) + \frac{1}{m} \sum_{i=1}^m (K_{x_i}, (f - f^o_{T,y}))^2_K + \gamma \|f - f^o_{T,y}\|_K^2$. Then for $\|f - f^o_{T,y}\|_K \leq t$, we get

$$|\Phi_{T,y}(f) - \Phi_{T,y}(f^o_{T,y})| \leq \frac{1}{m} \sum_{i=1}^m \|K_{x_i}\|_K^2 \|f - f^o_{T,y}\|_K^2 + \gamma \|f - f^o_{T,y}\|_K^2 \leq (s_K^2 + \gamma)t^2.$$
The upper bound on $d_{T,y,n,z}$ and the expression of $\Delta_{T,y,n,z}$ follow by inspection of the proof of Makovoz’s improvement over Maurey-Jones-Barron’s lemma, given in theorem 1 in Makovoz (1996) restated in terms of $G$-variation.

(ii) By the definition of $f_n$ and Kůrková and Sanguineti (2005b, prop. 4.1), we have

$$\|f_n - f^0_{T,y}\|^2_K \leq \frac{\alpha_{T,y}(d_{T,y,n,z}) + \varepsilon_n}{\gamma},$$

which, together with case i, implies the estimate, equation 4.2.

Theorem 2 gives the upper bound,

$$\inf_{f \in \text{span}_{G_K}} \Phi_{T,y}(f) - \Phi_{T,y}(f^0_{T,y}) \leq \frac{2(s^2_K + \gamma)(\xi_{\lfloor n/2\rfloor}(G_K)\|f^0_{T,y}\|_{G_K})^2}{n},$$  

(4.3)

which improves the estimate

$$\inf_{f \in \text{span}_{G_K}} \Phi_{T,y}(f) - \Phi_{T,y}(f^0_{T,y}) \leq A/\sqrt{n} + B/n,$$  

(4.4)

from Kůrková and Sanguineti (2005b), where $A$ and $B$ are positive constants defined there. The improvement is twofold. The first-order optimality condition exploited in the proof of theorem 5(i) allows improving the estimate of the modulus of continuity of $\Phi_{T,y}$ at $f^0_{T,y}$ (the linear term in $t$, which determines the term $A/\sqrt{n}$ in equation 4.4, is removed). Moreover, the term $\xi_{\lfloor n/2\rfloor}(G_K)$, which by definition is a nonincreasing function of $n$, may provide an overall rate better than $1/n$ in equation 4.3. For instance, if for some $l < m/2$ the elements of $G_K$ are well approximated by the centroids of $l$ clusters, then the term $\xi_l(G_K)$ is small, and so the upper bound on $\inf_{f \in \text{span}_{G_K}} \Phi_{T,y}(f) - \Phi_{T,y}(f^0_{T,y})$ provided by theorem 5(i) may be much smaller than $(1/n)(2(s^2_K + \gamma)\|f^0_{T,y}\|_{G_K}^2)$. This case is of particular interest when there are many similar samples and a small number of significant clusters; of course, this depends on the properties of the probability measure $\rho$.

**Remark 1.** Let $\bar{\Delta}_{T,y} = s^2_K\|f^0_{T,y}\|_{G_K}^2 - \|f^0_{T,y}\|_K^2$. By proposition 5.3(iii) of Kůrková and Sanguineti (2005b), $\bar{\Delta}_{T,y} \leq \frac{(s^2_K - \lambda)\|y\|_2^2}{(\gamma + \lambda\min)^2}$, where $\lambda\min$ is the minimum eigenvalue of $K$. Exploiting Maurey-Jones-Barron’s lemma (see Kůrková & Sanguineti, 2005a) instead of its improvement by theorem 1
of Makovoz (1996), used in the proof of theorem 5 (i), we get the worse but simpler upper bound $d_{T,y,n,z} \leq \sqrt{\frac{\Delta_{x,y}}{n}}$.

**Remark 2.** Bounds with a similar structure as equation 4.3 were derived in Zhang (2002a, 2002b, 2003) for Tikhonov regularization and, more generally, convex optimization problems whose domains are convex hulls of certain sets of functions. The following are some differences between the estimate, equation 4.4, and such results.

- The estimates derived in Zhang (2002a, 2002b, 2003) are of order $1/n$ and have a constructive nature, in the sense that algorithms are described to achieve them (for some particular cases; in Zhang, 2002a, geometric upper bounds are also derived; however, as remarked in Zhang, 2002a, they depend on quantities that may be ill behaved).
- Our bounds are nonconstructive, but in various cases, they are better than those given in Zhang (2002a, 2002b, 2003). This happens thanks to the term $\xi \lfloor \frac{n}{2} \rfloor (G_Kx)$ in equation 4.3, which takes into account properties of the set $G_K$. To fix ideas, let us consider the particular case of a Gram matrix $K[x]$ with only $l < m/2$ distinct columns (i.e., there are $m - l$ repeated samples). Then equation 4.3 with $n = 2l$ gives,

$$\inf_{f \in \text{span}_n G_K} \Phi_{T,y}(f) - \Phi_{T,y}(f_{T,y}^0) \leq 0.$$  \hspace{1cm} (4.5)

So, $\inf_{f \in \text{span}_n G_K} \Phi_{T,y}(f) = \Phi_{T,y}(f_{T,y}^0)$, whereas corollary 1 in Zhang (2002a) provides

$$\inf_{f \in \text{span}_n G_K} \Phi_{T,y}(f) - \Phi_{T,y}(f_{T,y}^0) \leq \Phi_{T,y}(0) \left(\frac{M^2 + 2\gamma}{(2\gamma)^2} \cdot 2l + (M^2 + 2\gamma)^2\right).$$  \hspace{1cm} (4.6)

where $M = \max_i \|Kx_i\|_K$.

- Differently from Zhang (2002a, 2002b, 2003), theorem 2(ii) also gives an upper bound on the quantity $\|f_n - f_{T,y}^0\|_K^2$, where $f_n$ is any $\varepsilon_n$-near minimum point of $(\text{span}_n G_K, \Phi_{T,y})$.

- The upper bounds of order $1/n$ provided in Zhang (2002a, 2002b, 2003) can be applied to the case of Tikhonov regularization but not to the cases of Ivanov, Miller, and Phillips regularizations. For instance, lemma 3.1 in Zhang, 2002b requires that the domain of the functional $\Phi$ to be minimized is the convex hull conv($S$) of a set $S$ of functions. In order to apply this lemma to Tikhonov regularization, one considers the functional $\Phi_{T,y}$ on conv($S$), where $S = \{f_{T,y}^0 \mid \|G_Kx\|_{G_K} \cup -f_{T,y}^0 \mid_{G_K} \}$. For Ivanov regularization, one should consider the functional $\Phi_{T,r}$ on conv($S$) $\cap B_r(\|\cdot\|_K)$, where $S = \{f_{T,r}^0 \mid G_Kx \cup -f_{T,r}^0 \mid_{G_K} \}$. However, in general conv($S$) $\not\subseteq$

---

2A refinement in the proof technique of Makovoz (1996, theorem 1) is required in order to replace $2l$ by $l$ in equation 4.5.
\(B_r(\| \cdot \| K)\), whereas the proof of Zhang’s lemma 3.1 works when the whole \(\text{conv}(S)\) is contained in the domain of the functional to be minimized.

We conclude this section by showing that under mild conditions, our approach provides SLT bounds that control, with high probability and uniformly over the family of functions within which we search for suboptimal solutions, the difference between the expected and the empirical errors (no estimates of this kind were given in Kurková and Sanguineti (2005b), Zhang (2002a, 2002b, 2003).

Suppose that an a priori upper bound \(\theta > 0\) is available on the absolute value of the output \(y\). Then \(\Phi_{T, \gamma}(0) = \mathcal{E}_z(0) = \frac{1}{m} \sum_{i=1}^{m} y_i^2 \leq \theta^2\). So as \(\Phi_{T, \gamma}\) is the sum of two nonnegative functionals, both \(f_{oT, \gamma}\) and any reasonable suboptimal solution \(f_n \in \text{span}_n G_K\) are such that \(\| f_{oT, \gamma} \| K \leq \theta \sqrt{\gamma}\) and \(\| f_n \| K \leq \theta \sqrt{\gamma}\) (it is not reasonable to consider a suboptimal solution \(f_n \in \text{span}_n G_K\) with \(\| f_n \| K > \theta \sqrt{\gamma}\), because in such a case, one would have \(\Phi_{T, \gamma}(f_n) \geq \Phi_{T, \gamma}(0)\), so \(0 \in \text{span}_n G_K\) would be a better suboptimal solution). Thus, if \(X\) is compact and the kernel \(K\) is continuous, theorem 7 in the appendix (with \(R_1 = \theta / \sqrt{\gamma}\) and \(R_2 = \theta \max\{1/s_K, 1/\sqrt{\gamma}\} = \nu_T(\gamma)\)) gives for every \(\gamma > 0\), every \(m \in \mathbb{N}\), and every \(\tau \in (0, 1)\),

\[
P\left\{ \sup_{f \in \text{B}_{\theta / \sqrt{\gamma}}(\| \cdot \| K)} (\mathcal{E}(f) - \mathcal{E}_z(f)) \leq \frac{16s_K^2 \nu_T(\gamma)}{\sqrt{m}} (\theta / \sqrt{\gamma} + \nu_T(\gamma)) \right. \\
+ 4(s_K \nu_T(\gamma))^2 \sqrt{\frac{\ln(2/\tau)}{2m}} \right\} \geq 1 - \tau, \quad (4.7)
\]

where \(P\) denotes the a priori probability with respect to the random draw of the i.i.d. data sample \(z\). Let \(P_{z, n} f_{oT, \gamma}\) be any best approximation of \(f_{oT, \gamma}\) on \(\text{span}_n G_K\), (such a best approximation exists, since \(\text{span}_n G_K\) is a finite union of convex and closed nonempty sets in a Hilbert space, and for each of them, there is a best approximation of \(f_{oT, \gamma}\) (Limaye, 1996, theorem 3.5)). By definition, \(\| f_{oT, \gamma} \| K = d_{T, \gamma, n, z}\), and by geometric properties of the set \(\text{span}_n G_K\), one has \(\| P_{z, n} f_{oT, \gamma} \| K \leq \| f_{oT, \gamma} \| K \leq \theta / \sqrt{\gamma}\). This shows that the bound 4.3 and the estimates from theorem 2 still hold, together with equation 4.7, if one replaces the constraint \(f \in \text{span}_n G_K\) with \(f \in \text{span}_n G_K \cap \text{B}_{\theta / \sqrt{\gamma}}(\| \cdot \| K)\) (the only change required in the proof of theorem 2(i) consists in exploiting the equality \(\| f_{T, \gamma} \| K = d_{T, \gamma, n, z}\).

Remark 3. The discussion above (which, for simplicity, has been limited to a fixed value of \(\gamma\)) can be improved in two ways. First, one can apply the structural risk minimization principle (Vapnik, 1998) in a form similar to the one in (Shawe-Taylor, Bartlett, Williamson, & Anthony, 1998, theorem 2.1) restated for regression problems and in terms of Rademacher’s
complexity instead of the VC dimension. In such a way, by fixing two sequences \( \{ \gamma_j > 0 \} \) and \( \{ a_j > 0 \} \) such that \( \sum_{j=1}^{+\infty} a_j = 1 \) and applying the union-bound technique, instead of equation 4.7, one gets.\(^3\)

\[
\mathbb{P}\left\{ \forall j \in \mathbb{N} \forall f \in B_{\theta/\sqrt{\gamma_j}}(\| \cdot \|_K) : (\mathcal{E}(f) - \mathcal{E}_z(f)) \leq \frac{16s^2_K v_T(\gamma_j)}{\sqrt{m}} (\theta / \sqrt{\gamma_j} + v_T(\gamma_j)) + 4(s_K v_T(\gamma_j))^2 \sqrt{\frac{\ln(2/(a_j \tau))}{2m}} \right\} \\
\geq 1 - \tau.
\] (4.8)

This suggests choosing (a posteriori) a value of \( \gamma_j \) in the sequence, which minimizes an upper bound on

\[
\inf_{f \in \text{span}_K G_K \cap B_{\theta/\sqrt{\gamma_j}}(\| \cdot \|_K)} \mathcal{E}_z(f) + \frac{16s^2_K v_T(\gamma_j)}{\sqrt{m}} (\theta / \sqrt{\gamma_j} + v_T(\gamma_j)) + 4(s_K v_T(\gamma_j))^2 \sqrt{\frac{\ln(2/(a_j \tau))}{2m}}.
\]

or (taking into account that \( \mathcal{E}_z(f) \leq \Phi_{T,\gamma_j}(f) \)) on

\[
\inf_{f \in \text{span}_K G_K \cap B_{\theta/\sqrt{\gamma_j}}(\| \cdot \|_K)} \Phi_{T,\gamma_j}(f) + \frac{16s^2_K v_T(\gamma_j)}{\sqrt{m}} (\theta / \sqrt{\gamma_j} + v_T(\gamma_j)) + 4(s_K v_T(\gamma_j))^2 \sqrt{\frac{\ln(2/(a_j \tau))}{2m}}.
\]

Second, one can improve the analysis starting from a data-dependent SLT bound like theorem 6 instead of the data-independent one provided by theorem 7, which does not take into account any information available from the data (e.g., in the proof of theorem 7, we have used the data-independent bound \( \int_X K(t, t) \rho_X(t) \leq s^2_K \), but a better estimate may be available from the data).

---

\(^3\)Suppose that the sequence \( \{ \gamma_j \} \) is decreasing and \( \lim_{j \to +\infty} \gamma_j = 0 \). Then for every \( f \in \mathcal{H}_K(X) \), one can apply the bound inside the braces of equation 4.8, taking \( \gamma \) equal to the largest \( \gamma_j \) in the sequence such that \( f \in B_{\theta/\sqrt{\gamma_j}}(\| \cdot \|_K) \). Extensions of equation 4.8, holding uniformly for all values of \( \gamma \) in a given interval may be obtained by following the technique used in Bousquet and Elisseeff (2002, proof of theorem 18).
5 Estimates for Ivanov-Regularized Learning

For Ivanov regularization, estimates of the accuracy of suboptimal solutions are given in the following theorem (recall that $B_r(\|\cdot\|_K)$ is the hypothesis set in Ivanov regularization).

**Theorem 3.** Let $X$ be a nonempty set, $K : X \times X \to \mathbb{R}$ a psd kernel with $s_K = \sup_{x \in X} \sqrt{K(x, x)} < +\infty$, $m$ a positive integer, $z$ a data sample of size $m$, $|y|_{\text{max}} = \max\{|y_i| : i = 1, \ldots, m\}$, $0 < r \leq \| f_i^0 \|_K$, and $\Delta_{1, r, n, z} = 2(\xi_{[n/2]}(G_{K_x})\|f_i^0\|_G_{K_x})^2$. Then for every positive integer $n < m$ the following hold:

i. Let $\alpha_{1, r}(t)$ be the modulus of continuity of $\Phi_{1, r}$ at $f_i^0$, and

$$d_{1, r, n, z} = \inf_{g \in \text{span}_n G_{K_x}} \| f_i^0 - g \|_K.$$ 

Then

$$\inf_{f \in B_r(\|\cdot\|_K) \text{span}_n G_{K_x}} \Phi_{1, r}(f) - \Phi_{1, r}(f_i^0) \leq \alpha_{1, r}(d_{1, r, n, z}),$$

(5.1)

where $\alpha_{1, r}(t) = a_1 t + a_2 t^2, a_1 = 2(r s_K^2 + |y|_{\text{max}} s_K)$, $a_2 = s_K^2$, and $d_{1, r, n, z} \leq \sqrt{\Delta_{1, r, n, z}}$.

ii. If one also has $r \leq \frac{\sqrt{\lambda_{\text{min}}} y_{\text{max}}}{\lambda_{\text{max}} - \lambda_{\text{min}}}$, where $\lambda_{\text{min}}$ and $\lambda_{\text{max}}$ are the minimum and maximum eigenvalues of $K[x]$ (resp.), then $\| f_i^0 \|_{G_{K_x}} \leq \frac{m\|y\|_{\text{max}}}{\sqrt{\lambda_{\text{min}}} y_{\text{max}} - \sqrt{\lambda_{\text{min}}} (\lambda_{\text{max}} - \lambda_{\text{min}})^2}$.

iii. Let $\varepsilon_n > 0$ and $f_n$ be any $\varepsilon_n$-near minimum point of $(B_r(\|\cdot\|_K) \cap \text{span}_n G_{K_x}, \Phi_{1, r})$. Then

$$\| f_n - f_i^0 \|_K^2 \leq \frac{1}{\gamma(r)} \left( \alpha_{1, r} \left( \sqrt{\frac{\Delta_{1, r, n, z}}{n}} \right) + \varepsilon_n \right) + \left( \| f_n \|_K^2 - r^2 \right).$$

iv. For every $n > m$, let $\{\varepsilon_{n, k}\}$ be a sequence of positive reals indexed by $k$, such that $\lim_{k \to +\infty} \varepsilon_{n, k} = 0$ and $\{f_{n, k}\}$ a sequence of $\varepsilon_{n, k}$-near minimum points of $(B_r(\|\cdot\|_K) \cap \text{span}_n G_{K_x}, \Phi_{1, r})$. Then

$$\limsup_{k \to +\infty} \| f_{n, k} - f_i^0 \|_K^2 \leq \alpha_{1, r} \left( \frac{\Delta_{1, r, n, z}}{n} \right)^{mr} \frac{\sqrt{\lambda_{\text{min}}} y_{\text{max}}}{\sqrt{\lambda_{\text{min}}} y_{\text{max}} - r \lambda_{\text{max}}}.$$

**Proof.** (i) Let $\mathcal{P}_{z, n} f_i^0$ be any best approximation of $f_i^0$ on $\text{span}_n G_{K_x}$. Then $\| f_i^0 - \mathcal{P}_{z, n} f_i^0 \|_K = d_{1, r, n, z}$ and $\| \mathcal{P}_{z, n} f_i^0 \|_K \leq \| f_i^0 \|_K$, so $\mathcal{P}_{z, n} f_i^0 \in B_r(\|\cdot\|_K) \cap \text{span}_n G_{K_x}$. Hence, equation 5.1 follows by the definition of the modulus of continuity.
Let us derive the upper bound on the modulus of continuity $\alpha_{l,r}$. For every $f, g \in \mathcal{H}_K(X)$,

$$|\Phi_{l,r}(f) - \Phi_{l,r}(g)| = \left| \frac{1}{m} \sum_{i=1}^{m} [(f(x_i) - y_i)^2 - (g(x_i) - y_i)^2] \right|$$

$$= \left| \frac{1}{m} \sum_{i=1}^{m} (f(x_i) - g(x_i))(f(x_i) + g(x_i) - 2y_i) \right|$$

$$\leq \sup_{x \in X} |f(x) - g(x)| \left( \sup_{x \in X} |f(x) + g(x)| + 2|y|_{\max} \right).$$

By the reproducing property 2.1, for every $x \in X$, we get $|f(x) - g(x)| = |(f - g, K_x)\| \leq \|f - g\|_K \sqrt{K(x, x)} \leq s_K \|f - g\|_K$, so

$$|\Phi_{l,r}(f) - \Phi_{l,r}(g)| \leq s_K \|f - g\|_K (s_K \|f + g\|_K + 2|y|_{\max}).$$

Let $\|f - g\|_K \leq t$. Then $\|g\|_K \leq \|f\|_K + t$, which, together with $\|f + g\|_K \leq \|f\|_K + \|g\|_K$, implies

$$|\Phi_{l,r}(f) - \Phi_{l,r}(g)| \leq 2(\|f\|_K s_K^2 + |y|_{\max} s_K) t + s_K^2 t^2. \quad (5.2)$$

For $0 < r \leq \|f_x^1\|_K$, by proposition 2(i), we have $f_{T,y(r)}^0 = f_{l,r}^0$, and so $\|f_{T,y(r)}^0\|_K = \|f_{l,r}^0\|_K = r$. Hence, $\alpha_{l,r}(t) \leq a_1 t + a_2 t^2$, where $a_1 = 2(rs_K^2 + |y|_{\max} s_K)$ and $a_2 = s_K^2$.

The upper bound on $d_{l,r,n,z}$ and the expression of $\Delta_{l,r,n,z}$ follow as in the proof of theorem 2(i).

(ii) As $0 < r \leq \|f_x^2\|_K$, by proposition 5.3(i) in Kúrková and Sanguineti (2005b) and the equivalence between Tikhonov and Ivanov regularizations (see proposition 2), we get

$$\|f_{l,r}^0\|_{G_{K_x}} = \|f_{T,y(r)}^0\|_{G_{K_x}} \leq \frac{\sqrt{m} \|y\|_2}{\gamma(r)m + \lambda_{\min}}.$$ 

Then by proposition 3 (i), we get

$$\frac{\sqrt{m} \|y\|_2}{\gamma(r)m + \lambda_{\min}} \leq \frac{\sqrt{m} \|y\|_2}{\gamma(r) \max m \cdot \lambda_{\min}},$$

where the denominator is positive as by assumption $r < \sqrt{\lambda_{\min}} \|y\|_2 \lambda_{\max}$. This implies the upper bound on $\|f_{l,r}^0\|_{G_{K_x}}^2$ given in the statement.

(iii) By proposition 5.1(iii) of Kúrková and Sanguineti (2005b) and the equivalence between Tikhonov and Ivanov regularizations (see
proposition 2), we get
\[ \| f_n - f^o_{I,r} \|_K^2 = \| f_n - f^o_{I,r}(r) \|_K^2 \leq \frac{\Phi_{T,r}(f_n) - \Phi_{T,r}(f^o_{I,r}) + \varepsilon_n}{\gamma(r)} \]
\[ = \frac{\Phi_{I,r}(f_n) - \Phi_{I,r}(f^o_{I,r}) + \varepsilon_n}{\gamma(r)} + \left( \| f_n \|_K^2 - r^2 \right), \]
which, combined with case i, gives case iii.

(iv) Combining case iii with \( \| f_{n,k} \|_K \leq r \) and proposition 3, we get
\[ \limsup_{k \to +\infty} \left\| f_{n,k} - f^o_{I,r} \right\|_K^2 \leq \frac{1}{\gamma(r)} \left( \alpha_{I,r} \left( \sqrt{\frac{\Delta_{I,r,n,z}}{n}} \right) \right) \]
\[ + \limsup_{k \to +\infty} \left( \| f_{n,k} \|_K^2 - r^2 \right) \]
\[ \leq \frac{1}{\gamma(r)} \left( \alpha_{I,r} \left( \sqrt{\frac{\Delta_{I,r,n,z}}{n}} \right) \right) \]
\[ \leq \left( \alpha_{I,r} \left( \sqrt{\frac{\Delta_{I,r,n,z}}{n}} \right) \right) \frac{mr}{\sqrt{\lambda_{\min} \| y \|_2 - r \lambda_{\max}}}. \]

Theorem 3(i) provides an estimate of the form
\[ \inf_{f \in B_r(\| \cdot \|_K) \cap \text{span}_n G_K} \Phi_{I,r}(f) - \Phi_{I,r}(f^o_{I,r}) \]
\[ \leq A \frac{\xi_{[n/2]}(G_K)}{\sqrt{n}} + B \frac{(\xi_{[n/2]}(G_K))^2}{n}, \quad (5.3) \]
for suitable constants \( A, B > 0 \) (a worse but simpler upper bound on \( d_{I,r,n,z} \) can be obtained proceeding as described in remark 1).

Now suppose that for a fixed \( r \), the data generation model is such that
\[ \mathbf{P}\{0 < r \leq \| f^\dagger \|_K \} \geq 1 - \mu_1(r, m), \quad (5.4) \]
for some function \( \mu_1 : (0, +\infty) \times \mathbb{N} \to [0, 1) \). As done in section 4 for Tikhonov regularization, if one has an a priori upper bound \( \theta > 0 \) on the absolute value of the output \( y \), \( X \) is compact, and the kernel \( K \) is continuous, then theorem 7 in the appendix (with \( R_1 = r \) and \( R_2 = \max\{ \frac{\theta}{r_k}, r \} = \nu_1(r) \)), combined with equation 5.4 and the union-bound technique, gives, for
every \( r > 0 \), every \( m \in \mathbb{N} \), and every \( \tau \in (0, 1) \),
\[
P \left\{ 0 < r \leq \| f^*_z \|_K \quad \& \quad \sup_{f \in B_r(\| \cdot \|_K)} (\mathcal{E}(f) - \mathcal{E}_z(f)) \leq \frac{16s^2_K v_1(r) (r + v_1(r))}{\sqrt{m}} \right\} \geq 1 - \tau - \mu_1(r, m). \tag{5.5}
\]

6 Estimates for Phillips- and Miller-Regularized Learning

For Phillips and Miller regularizations, the proof technique of theorem 3(i) cannot be applied for the following reason (a similar remark holds for the proof technique of theorem 2(i)). In the proof of theorem 3(i), we have shown that \( P_{z,n} f^o_{l,r} \) belongs to the hypothesis set \( B_r(\| \cdot \|_K) \) of Ivanov regularization. However, in general, \( P_{z,n} f^o_{l,r} \) and \( P_{z,n} f^o_{M,r,\eta} \) do not belong to the hypothesis sets \( F_{z,\eta} \) and \( G_{z,r,\eta} \) of Phillips and Miller regularizations, respectively.

A different way to obtain rates of approximate optimization in the Phillips and Miller regularizations is suggested by theorem 4.2(i) in Kurková and Sanguineti (2005a), which, however, requires that \( 0 \) is in the topological interior of \( F_{z,\eta} \) and \( G_{z,r,\eta} \), respectively, whereas in general, even \( 0 \notin F_{z,\eta} \) and \( 0 \notin G_{z,r,\eta} \). Note that instead, all the other hypotheses of theorem 4.2(i) of Kurková and Sanguineti (2005a) are satisfied. A value of \( \eta \) sufficiently large would allow guaranteeing \( 0 \in \text{int}(F_{z,\eta}) \) and \( 0 \in \text{int}(G_{z,r,\eta}) \); however, such a large value might determine an undesired tolerance on the data misfit of the corresponding regularized solutions. For these reasons, to estimate the accuracy of suboptimal solutions to Phillips and Miller regularizations, we use a different approach, which also has the advantage of being applicable to all four regularization methods. This approach can be summarized in the following steps:

1. If assumption 1 holds, then by theorem 1 the solutions to Tikhonov, Ivanov, Phillips, and Miller regularizations belong to the family \( F_z \) of functions, defined as
\[
F_z = \left\{ f^\alpha_z = \sum_{i=1}^{m} c_{\alpha,i} K_{x_i} : \alpha > 0 \quad \text{and} \quad c_{\alpha} = (c_{\alpha,1}, \ldots, c_{\alpha,m})^T \quad \text{is the unique solution of the well-posed linear system} \quad (\alpha \ m I + K[x])c_{\alpha} = y \right\}.
\]
2. For every \( f^\alpha_z \in F_z \), we shall derive upper bounds on \( \| f^\alpha_z \|_K \) and \( \mathcal{E}_z(f^\alpha_z) \).
3. We shall combine the bounds from step 2 with an upper bound on the approximation error \( \|f_z^\alpha - P_{z,n}f_z^\alpha\|_K \), where \( P_{z,n}f_z^\alpha \) is any best approximation of \( f_z^\alpha \) on \( \text{span}_n G_{Kz} \).

4. For each regularization method, we shall exploit the bounds obtained in step 3 to find a set of values of \( \alpha \) for which the function \( P_{z,n}f_z^\alpha \) belongs to the hypothesis set of the corresponding regularization problem, and we choose the value of \( \alpha \) that gives the smallest upper bound on the corresponding functional.

The next proposition provides the bounds on \( \|f_z^\alpha\|_K \) and \( \mathcal{E}_z(f_z^\alpha) \) required in step 2.

**Proposition 7.** Let \( X \) be a nonempty set, \( K : X \times X \rightarrow \mathbb{R} \) a psd kernel, \( m \) a positive integer, \( z \) a data sample of size \( m \), and \( \lambda_{\min} \) and \( \lambda_{\max} \) the minimum and maximum eigenvalues of \( K[x] \), respectively. Then for every \( f_z^\alpha \in \mathcal{F}_z \), the following hold:

\[
\begin{align*}
&\text{i. } \frac{\sqrt{\lambda_{\min}}}{a_m + \lambda_{\max}} \| y \|_2 \leq \| f_z^\alpha \|_K \leq \frac{\sqrt{\lambda_{\max}}}{a_m + \lambda_{\min}} \| y \|_2, \\
&\text{ii. } \frac{a^2 m}{\lambda_{\min} + \lambda_{\max}} \| y \|_2^2 \leq \mathcal{E}_z(f_z^\alpha) \leq \frac{a^2 m}{\lambda_{\max} + \lambda_{\min}} \| y \|_2^2.
\end{align*}
\]

**Proof.** (i) We proceed as in the proof of proposition 3(i) (better estimates, with a more complex form, can be obtained by proceeding as in the proofs of proposition 3(ii) or proposition 4). As \( f_z^\alpha = \sum_{i=1}^m c_i K_x \), by the reproducing property 2.1, we get \( \|f_z^\alpha\|_K = \sqrt{c_i^2 K[x] c_i} \). Since \( c_i = (\alpha m I + K[x])^{-1} y \), we have \( \|f_z^\alpha\|_K \leq \sqrt{\lambda_{\max}} \|c_i\|_2 \leq \frac{\sqrt{\lambda_{\max}}}{a_m + \lambda_{\min}} \|y\|_2 \) and \( \|f_z^\alpha\|_K \geq \sqrt{\lambda_{\min}} \|c_i\|_2 \geq \frac{\sqrt{\lambda_{\min}}}{a_m + \lambda_{\max}} \|y\|_2 \).

(ii) By the theory of Tikhonov regularization (see Bertero, 1989, and Engl et al., 2000), we get

\[
\mathcal{E}_z(f_z^\alpha) = \frac{1}{m} \| L_x L_x^* (L_x L_x^* + \alpha I)^{-1} y - y \|_2^2 = \frac{a^2}{m} \| (L_x L_x^* + \alpha I)^{-1} y \|_2^2,
\]

where \( L_x : \mathcal{H}_K(X) \rightarrow \mathbb{R}^m \) is the operator defined for every \( f \in \mathcal{H}_K(X) \) as \( L_x(f) = (\frac{f(x_1)}{\sqrt{m}}, \ldots, \frac{f(x_m)}{\sqrt{m}}) \). By proposition 3.1(iv) of Kůrková (2004), the operator \( L_x L_x^* : \mathbb{R}^m \rightarrow \mathbb{R}^m \) can be represented in the canonical basis of \( \mathbb{R}^m \) by the matrix \( \frac{1}{m} K[x] \), so we get

\[
\mathcal{E}_z(f_z^\alpha) = a^2 m \| (K[x] + \alpha m I)^{-1} y \|_2^2.
\]

from which the statement easily follows proceeding as in the proof of (i).

The next proposition provides the bounds on \( \|P_{z,n}f_z^\alpha\|_K \) and \( \mathcal{E}_z(P_{z,n}f_z^\alpha) \) required by step 3 by combining the estimates on \( \|f_z^\alpha\|_K \) and \( \mathcal{E}_z(f_z^\alpha) \) from proposition 7, with an upper bound on \( \|f_z^\alpha - P_{z,n}f_z^\alpha\|_K \):
Proposition 8. Let $X$ be a nonempty set, $K: X \times X \rightarrow \mathbb{R}$ a psd kernel with $s_K = \sup_{x \in X} \sqrt{K(x, x)} < +\infty$, $m$ a positive integer, $Z$ a data sample of size $m$, and $\lambda_{\min}$, $\lambda_{\max}$ the minimum and maximum eigenvalues of $K[x]$, respectively. For $f^a \in \mathcal{F}$, let $\mathcal{P}_{\alpha,n} f^a_z$ be a best approximation of $f^a$ on $\mathcal{G}_{X\alpha}$, $u_{\alpha} = \sqrt{s_K^2 \|f^a\|^2_{\mathcal{G}_{X\alpha}} - \|f^a\|^2_{\mathcal{F}}}$, and $u = \sqrt{s_K^2 m - \lambda_{\min}}$. Then:

i. $\|\mathcal{P}_{\alpha,n} f^a_z\|_K \leq \|f^a\|_K + \frac{u_{\alpha}}{\sqrt{n}}$.

ii. $\|\mathcal{P}_{\alpha,n} f^a_z\|_K \leq \frac{\sqrt{\lambda_{\max} \|f\|^2}}{\eta (am + \lambda_{\min})} + \frac{\|f\|^2}{\eta (am + \lambda_{\min})} \frac{u}{\sqrt{n}}$.

iii. $\mathcal{E}_z(\mathcal{P}_{\alpha,n} f^a_z) \leq \mathcal{E}_z(f^a) + \left(2 \|f^a\|_K s_K^2 + \|y\|_{\max} s_K \right) \frac{u_{\alpha}}{\sqrt{n}} + s_K^2 \frac{u^2}{n}$.

iv. $\mathcal{E}_z(\mathcal{P}_{\alpha,n} f^a_z) \leq \frac{\alpha \eta m \|f\|^2}{(am + \lambda_{\min})^2} + \frac{s_K \|f\|^2}{(am + \lambda_{\min}) \sqrt{\lambda_{\max} + \|y\|_{\max}}} \frac{u}{\sqrt{n}} + \frac{s_K^2 \|f\|^2}{(am + \lambda_{\min})^2} \frac{u^2}{n}$.

Proof. Let $d_{a,n,z} = \inf_{g \in \text{span}_{\alpha \mathcal{G}_{X\alpha}}} \|f^a_z - g\|_K = \|\mathcal{P}_{\alpha,n} f^a_z - f^a\|_K$.

Case i follows by combining $\|\mathcal{P}_{\alpha,n} f^a_z\|_K \leq \|f^a\|_K + d_{a,n,z}$ with the upper bound on $d_{a,n,z}$ from theorem 3.1(i) of Kůrková and Sanguineti (2005a) (better estimates, with a more complex expression, can be obtained exploiting the upper bound on $d_{a,n,z}$ from theorem 1 of Makovoz (1996)).

Case ii follows by combining case i with the upper bound on $\|f^a\|_K$ from proposition 7(i) and the upper bound $u_{\alpha}^2 = \frac{s_K^2 m - \lambda_{\min}}{(am + \lambda_{\min})^2}$ from proposition 5.3(iii) of Kůrková and Sanguineti (2005b).

For case iii, by equation 5.2, $|\mathcal{E}_z(\mathcal{P}_{\alpha,n} f^a_z) - \mathcal{E}_z(f^a)| \leq 2(\|f^a\|_K s_K^2 + \|y\|_{\max} s_K t + s_K^2 t^2$ for every $t \geq \|\mathcal{P}_{\alpha,n} f^a_z - f^a\|_K = d_{a,n,z}$. We conclude by combining this with the upper bound on $d_{a,n,z}$ from theorem 3.1(iii) of Kůrková and Sanguineti (2005a).

For case iv, it follows, combining case iii with the upper bounds on $\mathcal{E}_z(f^a)$, $\|f^a\|_K$, and $u_{\alpha}$ by proposition 7(ii) (Kůrková & Sanguineti, 2005b, propositions 5.3(ii), and 5.3(iii), respectively).

The choice between the bounds given in proposition 8(i) and 8(ii) and, similarly, the choice between those given in proposition 8(iii) and 8(iv) depends on the availability of sufficiently accurate expressions for $\|f^a\|_K$, $\mathcal{E}_z(f^a)$ and $\|f^a\|_{\mathcal{G}_{X\alpha}}$ as functions of $\alpha$.

Remark 4. It is worth remarking on the behavior for $\alpha \rightarrow 0$ and $\alpha \rightarrow +\infty$ of the bounds provided by proposition 8. For $\alpha \rightarrow 0$, one has $f^a_z \rightarrow f^a$ in the norm of the RKHS by the theory of Tikhonov regularization (Engl et al., 2000), $\mathcal{E}_z(f^a) \rightarrow \mathcal{E}_z(f^a)$ (as a consequence of the reproducing property 2.1), and $u_{\alpha}^2 = s_K^2 \|f^a\|^2_{\mathcal{G}_{X\alpha}} - \|f^a\|^2_{\mathcal{F}} \rightarrow s_K^2 \|f^a\|^2_{\mathcal{G}_{X\alpha}} - \|f^a\|^2_{\mathcal{F}}$ (since $G$-variation is a norm—Kůrková, 1997—span $\mathcal{G}_{X\alpha}$ is finite-dimensional, and all norms are equivalent in a finite-dimensional space—Bachmann & Narici, 2000, theorem 8.7). Thus, setting $u_{\alpha} = \sqrt{s_K^2 \|f^a\|^2_{\mathcal{G}_{X\alpha}} - \|f^a\|^2_{\mathcal{F}}}$, for
\( \alpha \to 0 \) proposition 8(i) and 8(iii) implies \( \lim_{\alpha \to 0} \| P_{z,n} f^\alpha_z \|_K \leq \| f^1_z \|_K + \frac{u_1}{\sqrt{n}} \) and \( \lim_{\alpha \to 0} E_z(P_{z,n} f^\alpha_z) \leq E_z(f^1_z) + (2 \| f^1_z \|_K s^2_K + |y|_{\max s_K}) \frac{u_1}{\sqrt{n}} + s^2_K \frac{u_2^2}{n} \). Similarly, for \( \alpha \to +\infty \), one has \( f^\alpha_z \to 0 \), \( E_z(f^\alpha_z) \to E_z(0) \), and \( u_2^\alpha = s^2_K \| f^\alpha_z \|^2_{G_{\alpha z}} - \| f^\alpha_z \|^2_K \to 0 \). So for \( \alpha \to +\infty \) proposition 8 (ii) and the reproducing property 2.1 imply \( \lim_{\alpha \to +\infty} \| P_{z,n} f^\alpha_z \|_K = 0 \) and \( \lim_{\alpha \to +\infty} E_z(P_{z,n} f^\alpha_z) = \frac{|y|_2^2}{m} = E_z(0) \). This analysis of the limit behavior of \( f^\alpha_z \) allows one to parameterize \( f^\alpha_z \) in terms of \( \alpha \in [0, +\infty] \).

Finally, step 4 is performed as follows for each regularization method. As done in sections 4 and 5, we assume that an a priori upper bound \( \theta > 0 \) on the absolute value of the output \( y \) is available, that \( X \) is compact, and that the kernel \( K \) is continuous. For simplicity of exposition, we suppose that the expressions of \( \| f^\alpha_z \|_K \) and \( E_z(f^\alpha_z) \) are available; when this is not the case, one can exploit, instead of the upper bounds given by proposition 8(i) and 8(iii), those from proposition 8(ii) and 8(iv), respectively. Finally, without loss of generality we suppose that the minima of equations 6.1, 6.3, 6.6, and 6.10 with respect to \( \alpha \) in the respective sets are achieved (otherwise, one considers \( \varepsilon \)-near minimum points).

### 6.1 Tikhonov Regularization

For \( \gamma > 0 \) and a positive integer \( n < m \), we search for \( \bar{\alpha} \in [0, +\infty] \) that minimizes

\[
E_z(f^\bar{\alpha}_z) + (2 \| f^\bar{\alpha}_z \|_K s^2_K + |y|_{\max s_K}) \frac{u_\alpha}{\sqrt{n}} + s^2_K \frac{u_2^\alpha}{n} + \gamma \left( \| f^\bar{\alpha}_z \|_K + \frac{u_\alpha}{\sqrt{n}} \right)^2.
\]  

By remark 4, for \( \alpha = +\infty \) we have

\[
E_z(f^{+\infty}_z) + (2 \| f^{+\infty}_z \|_K s^2_K + |y|_{\max s_K}) \frac{u_{+\infty}}{\sqrt{n}} + s^2_K \frac{u_2^{+\infty}}{n} + \\
+ \gamma \left( \| f^{+\infty}_z \|_K + \frac{u_{+\infty}}{\sqrt{n}} \right)^2 = E_z(0) \leq \theta^2.
\]

So \( \| P_{z,n} f^{\bar{\alpha}}_z \|_K \leq \| f^{\bar{\alpha}}_z \|_K \leq \frac{\theta}{\sqrt{\gamma}} \), and by proposition 8(i) and 8(iii), we get

\[
\inf_{f \in \text{span}_{G_{\bar{\alpha}}, \ell_2} B_{\bar{\alpha}, \gamma}(\| \cdot \|_K)} \Phi_{T,\gamma}(f) \leq \Phi_{T,\gamma}(P_{z,n} f^{\bar{\alpha}}_z)
\]

\[
\leq E_z(f^{\bar{\alpha}}_z) + (2 \| f^{\bar{\alpha}}_z \|_K s^2_K + |y|_{\max s_K}) \frac{u_\alpha}{\sqrt{n}} + \\
+ s^2_K \frac{u_2^\alpha}{n} + \gamma \left( \| f^{\bar{\alpha}}_z \|_K + \frac{u_\alpha}{\sqrt{n}} \right)^2.
\]  

(6.2)
Recall that for every $\gamma > 0$, every positive integer $m$, and every $\tau \in (0, 1)$, the SLT bound, equation 4.7 holds.

6.2 Ivanov Regularization. For $r > 0$ and a positive integer $n < m$, consider the set $I_r = \{ \alpha \in [0, +\infty] : \|f_\alpha^z\|_K + \frac{u_\alpha}{\sqrt{n}} \leq r \}$. By proposition 8(i), $\alpha \in I_r$ implies $\|P_z(n f_\alpha^z)\|_K \leq r$ and by remark 4, $I_r \neq \emptyset$ (indeed, $\alpha = +\infty \in I_r$). Then we search for $\bar{\alpha} \in I_r$ that minimizes

$$E_z(f_\alpha^z) + (2\|f_\alpha^z\|_K s_K^2 + |y|_{\max} s_K) \frac{u_\alpha}{\sqrt{n}} + s_K^2 \frac{u_\alpha^2}{n}.$$  

(6.3)

By proposition 8(iii), we get

$$\inf_{f \in B_r(||\cdot||_K) \cap \text{span}_n G_{Kz}} \Phi_{I,r}(f) \leq \Phi_{I,r}(P_z(n f_\alpha^z)) \leq E_z(f_\alpha^z) + (2\|f_\alpha^z\|_K s_K^2 + |y|_{\max} s_K) \frac{u_\alpha}{\sqrt{n}} + s_K^2 \frac{u_\alpha^2}{n}.$$  

(6.4)

Following the same arguments as at the end of section 5, for every $r > 0$, every $m \in \mathbb{N}$, and every $\tau \in (0, 1)$, we have the SLT bound

$$P\left\{ \sup_{f \in B_r(||\cdot||_K)} (E(f) - E_z(f)) \leq \frac{16s_K^2 v_l(r)}{\sqrt{m}}(r + v_l(r)) + 4(s_K v_l(r))^2 \sqrt{\frac{\ln(2/\tau)}{2m}} \right\} \geq 1 - \tau.$$  

(6.5)

6.3 Miller Regularization. Suppose that the data generation model is such that for $r > 0$, $\eta \geq 0$, and positive integers $m$ and $n < m$, the set

$$A_{z,r,\eta,n} = \{ \alpha \in [0, +\infty] : \|f_\alpha^z\|_K + \frac{u_\alpha}{\sqrt{n}} \leq r \quad \& \quad E_z(f_\alpha^z) + (2\|f_\alpha^z\|_K s_K^2 + |y|_{\max} s_K) \frac{u_\alpha}{\sqrt{n}} + s_K^2 \frac{u_\alpha^2}{n} \leq \eta^2 \}$$  

(6.5)

is nonempty with a priori probability at least $1 - \mu_M(r, \eta, m, n)$, for a suitable $[0,1)$-valued function $\mu_M(r, \eta, m, n)$ (the condition $A_{z,r,\eta,n} \neq \emptyset$ implies $G_{z,\eta,r} \neq \emptyset$). Then we search for $\bar{\alpha} \in A_{z,r,\eta,n}$ that minimizes

$$E_z(f_\alpha^z) + (2\|f_\alpha^z\|_K s_K^2 + |y|_{\max} s_K) \frac{u_\alpha}{\sqrt{n}} + s_K^2 \frac{u_\alpha^2}{n} + \left( \frac{\eta}{r} \right)^2 \left( \|f_\alpha^z\|_K + \frac{u_\alpha}{\sqrt{n}} \right)^2.$$  

(6.6)
By proposition 8(i) and 8(iii), we get

\[
\inf_{f \in G_{z,n} \cap \text{span}_n G_{K_x}} \Phi_{M,r,n}(f) \leq \Phi_{M,r,n}(P_{z,n} f^0_z) \\
\leq \mathcal{E}_z(f^0_z) + (2 \|f^0_z\|_{\mathcal{K}^2} + \|y\|_{\max} s_K) \frac{u_a}{\sqrt{n}} + s^2_K \frac{u^2_a}{n} \\
+ \left( \frac{\eta}{r} \right)^2 \left( \|f^0_z\|_{\mathcal{K}} + \frac{u_a}{\sqrt{n}} \right)^2.
\]

(6.7)

Following the same arguments used to derive equation 5.5, for every \( r > 0 \), every \( \eta \geq 0 \), every positive integer \( n, m \) with \( n < m \), and every \( \tau \in (0, 1) \), we have the SLT bound

\[
P\left\{ A_{z,r,\eta,n} \neq \emptyset \quad \& \quad \sup_{f \in B_r(\|\cdot\|_{\mathcal{K}})} (\mathcal{E}(f) - \mathcal{E}_z(f)) \leq \frac{16s^2_K v_I(r)}{\sqrt{m}} (r + v_I(r)) \\
+ 4(s_K v_I(r))^2 \sqrt{\ln(2/\tau)} \right\} \geq 1 - \tau - \mu_M(r, \eta, m, n).
\]

(6.8)

The estimates derived above for Tikhonov, Ivanov, and Miller regularizations can be refined by applying the techniques mentioned in remark 3 and making suitable assumptions on the data generation model. The estimates that we derive for Phillips regularization employ a refinement based on remark 3.

6.4 Phillips Regularization. This case requires a different approach from the previous ones. Indeed, in Phillips regularization, one has a given upper bound on the square of the empirical error \( \mathcal{E}_z \), and one aims to minimize \( \|\cdot\|_{\mathcal{K}}^2 \) under such a constraint. This can be interpreted as follows. Fix two sequences \( \{r_j > 0\} \) and \( \{a_j > 0\} \) such that \( \sum_{j=1}^{+\infty} a_j = 1 \). Then an SLT bound of the form

\[
P\left\{ \forall j \in \mathbb{N}, \forall f \in B_{r_j}(\|\cdot\|_{\mathcal{K}}) : (\mathcal{E}(f) - \mathcal{E}_z(f)) \leq \frac{16s^2_K v_I(r_j)}{\sqrt{m}} (r_j + v_I(r_j)) \\
+ 4(s_K v_I(r_j))^2 \sqrt{\ln(2/(a_j \tau))} \right\} \geq 1 - \tau
\]

(6.9)

can be easily derived (compare it with equations 4.8 and 5.5). Suppose that \( \{r_j\} \) is an increasing sequence, \( \lim_{j \to \infty} r_j = +\infty \) and \( \{a_j\} \) is a decreasing sequence. Note that the sequence \( \{v_I(r_j)\} = \{\max\{a_j, r_j\}\} \) is nondecreasing. Then, when looking for a suboptimal solution to problem \( P_{\eta} \) in \( F_{z,\eta} \cap \text{span}_n G_{K_x} \), the smallest upper bound on \( \mathcal{E}(f) - \mathcal{E}_z(f) \) in equation 6.9 is obtained for the minimum \( r_j \) in the sequence such that the set
F_{z,\eta} \cap \text{span}_{n} G_{K} \cap B_{r_{j}}(\| \cdot \|_{K}) \text{ is nonempty} \ (\text{one might search for the smallest upper bound on } \mathcal{E}(f) \text{ on the whole span}_{n} G_{K}, \text{ but it does not correspond to Phillips regularization in strict sense}).

For } \eta \geq 0 \text{ and a positive integer } n < m, \text{ consider the set } \alpha \in [0, +\infty] : E_{z}(f_{\alpha}^{\eta}) = \sum_{i=1}^{m} c_{i} K_{x_{i}}(\bar{x}) \leq \max_{i} s_{K}^{2} + \frac{\| y \|_{\max}}{\sqrt{n}} + \frac{u_{\alpha}}{\sqrt{n}} \leq \eta^{2} \}. \text{ By proposition 8(iii), } \alpha \in A_{\eta} \text{ implies } E_{z}(\mathcal{P}_{z,n} f_{\alpha}^{\eta}) \leq \eta^{2}, \text{ and by remark 4, if the kernel is } pd \text{ and } x \text{ has no repeated entries } x_{i}, \text{ then } A_{\eta} \neq \emptyset \ (\text{indeed, } \alpha = 0 \in A_{\eta}). \text{ Then we search for } \bar{\alpha} \in A_{\eta} \text{ that minimizes }

\| f_{z} \|_{K} + \frac{u_{\bar{\alpha}}}{\sqrt{n}}. \quad (6.10)

So by proposition 8(i), we get

\inf_{f \in F_{z,\eta} \cap \text{span}_{n} G_{K}} \Phi_{P,\eta}(f) \leq \Phi_{P,\eta}(\mathcal{P}_{z,n} f_{\eta}^{\bar{\alpha}}) \leq \| f_{\bar{\alpha}} \|_{K} + \frac{u_{\bar{\alpha}}}{\sqrt{n}}. \quad (6.11)

We conclude this section by considering suboptimal solutions in span_{n} G_{K}, \text{ for } n < m.

For the gaussian kernel, the next proposition shows that in general, it is not possible to express a function of the form } f = \sum_{i=1}^{m} c_{i} K_{x_{i}} \text{ as a finite linear combination of gaussians with centers not all equal to the } m \text{ data points } x_{i}, \ i = 1, \ldots, m, \text{ unless an approximation error is introduced. More generally, the proposition holds for every } pd \text{ kernel. Linear independence of gaussians with varying widths and centroids was proven in Kůrková and Neruda (1994).}

**Proposition 9.** Let } K \text{ be the gaussian kernel, } x_{i} \in \mathbb{R}^{d}, \ i = 1, \ldots, m, \text{ and } f = \sum_{i=1}^{m} c_{i} K_{x_{i}} \text{ such that at least one } c_{i} \text{ is different from 0. Then for every positive integer } k, \text{ every } \{ \hat{c}_{j} \in \mathbb{R} : j = 1, \ldots, k \}, \text{ and every set } \{ \hat{x}_{j} \in \mathbb{R}^{d} : j = 1, \ldots, k \}, \text{ if } \{ x_{i} : i = 1, \ldots, m \text{ and } c_{i} \neq 0 \} \nsubseteq \{ \hat{x}_{j} : j = 1, \ldots, k \} \text{ there exists } \bar{x} \in \{ x_{i} \} \cup \{ \hat{x}_{j} \} \text{ such that }

\sum_{i=1}^{m} c_{i} K_{x_{i}}(\bar{x}) \neq \sum_{j=1}^{k} \hat{c}_{j} K_{\hat{x}_{j}}(\bar{x}).

**Proof.** The statement is a consequence of the fact that for every positive integer } h, \text{ } h \text{ gaussians with equal widths but different centers form a set of linearly independent functions. Indeed, for any } h \text{-tuple of input data, the corresponding Gram matrix has full rank (Schölkopf & Smola, 2002, theorem 2.18).}

However, studying approximation error bounds achievable by families of functions of the form } \sum_{j=1}^{k} c_{j} K_{\hat{x}_{j}} \text{ is still of interest, especially when
$k = n \ll m$. Inspection of the proofs of our results shows that with the exception of proposition 6, they still hold if one considers $\text{span}_n G_K$ instead of $\text{span}_n G_{K_\alpha}$ as $\text{span}_n G_{K_\alpha} \subseteq \text{span}_n G_K$. Estimating the improvement obtained by using the larger hypothesis set $M \cap \text{span}_n G_K$ instead of $M \cap \text{span}_n G_{K_\alpha}$ may be a subject of future research.

7 On Algorithms for Sparse Suboptimal Solutions

Various algorithms have been proposed to find sparse suboptimal solutions to approximation and optimization problems. In the following, we briefly focus on those useful for the regularization techniques considered in this letter.

The context common to all such algorithms is the following. Given a (typically redundant) set $D$ of functions, called dictionary, which are elements of a finite- or infinite-dimensional Hilbert space $\mathcal{H}$, for a “small” positive integer $n$, one aims to find an accurate suboptimal solution $f_s^n$ from $\text{span}_n D$ to a function approximation problem or, more generally, to a functional optimization problem.

When no structure is imposed on the dictionary $D$, the problem of finding the best approximation of a function $f \in \mathcal{H}$ from $\text{span}_n D$ is NP-hard (Davis, Mallat, & Avellaneda, 1997). However, the problem may drastically simplify when the elements of the dictionary have a suitable structure. The simplest situation arises when they are orthogonal. The case of a dictionary with nearly orthogonal elements, that is, a dictionary with small coherence (Gribonval & Vandergheynst, 2006), stays halfway between these two extremes and provides a computationally tractable problem, for which constructive approximation results are available (Das & Kempe, 2008; Gilbert, Muthukrishnan, & Strauss, 2003; Gribonval & Vandergheynst, 2006; Tropp, 2004). As in our context $D = G_{K_\alpha}$, we may want to choose a kernel $K$ such that the dictionary $G_{K_\alpha}$ has small coherence. If this is not possible, then, as in Honeine et al. (2007), one may consider as dictionary a suitable subset of $G_{K_\alpha}$ with a small coherence.

In the remaining of this section, we discuss three families of algorithms to derive sparse suboptimal solutions.

7.1 Greedy Algorithms. Starting from an initial sparse suboptimal solution $f_s^n$ with a small $n$ (usually $n = 0$ and $f_s^0 = 0$), typically greedy algorithms obtain inductively an $(n + 1)$-term suboptimal solution $f_s^{n+1}$ as a linear combination of the $n$-term one $f_s^n$ and a new element from the dictionary. So a sequence of low-dimensional optimization problems has to be solved. Depending on how such problems are defined, different kinds of greedy algorithms are obtained (see, e.g., Zhang, 2002a, 2002b, 2003). These algorithms are particularly suitable to derive sparse suboptimal solutions to Tikhonov regularization; however, in some cases, their rates of approximate optimization are worse than our estimate 4.3, as discussed in
section 4. Moreover, the rates derived in Zhang (2002a, 2002b, 2003) do not apply to Ivanov, Miller, and Phillips regularizations (the possibility of extending Zhang, 2002b, lemma 3.1, to such regularizations may be a subject of future research; see also remark 2).

Remarkable properties were proven for matching pursuit and orthogonal matching pursuit; SLT bounds for their kernel versions, known as kernel matching pursuit (Vincent & Bengio, 2002), were derived in Hussain and Shawe-Taylor (2009, theorem 3). Given \( f \in \mathcal{H} \) and provided that the positive integer \( n \) and the dictionary \( D \) are suitably chosen, the \( n \)-term approximations found by these two algorithms are only a well-defined factor \( C(n) > 0 \) worse than the best approximation of \( f \) in terms of any \( n \) elements of \( D \) (see Gilbert et al., 2003, theorem 2.1; Tropp, 2004, theorem 2.6; Gribonval & Vandergheynst, 2006, featured theorem 3; Das & Kempe, 2008, theorem 3.5) for some values of \( C(n) \) and estimates on the number of the iterations. In the framework of section 6 and under the assumption that for a given \( \alpha > 0 \), the function \( f^\alpha \) is known,4 kernel versions of matching pursuit and orthogonal matching pursuit may be applied to find its sparse approximation \( f^s_{z,n} \). If \( \| f^\alpha_z - f^s_{z,n} \|_K \leq C(n)\| f^\alpha_z - P^\alpha_{z,n}f^\alpha_z \|_K \), then \( \| P^\alpha_{z,n}f^\alpha_z - f^s_{z,n} \|_K \leq (C(n) + 1)\| f^\alpha_z - P^\alpha_{z,n}f^\alpha_z \|_K \). Note that the results of section 6 provide upper bounds on \( \| f^\alpha_z - P^\alpha_{z,n}f^\alpha_z \|_K \). The differences \( |E_z(f^\alpha_z) - E_z(f^s_{z,n})| \) and \( |E_z(P^\alpha_{z,n}f^\alpha_z) - E_z(f^s_{z,n})| \) can be bounded from above in terms of the modulus of continuity of the functional \( \mathcal{E}_z \).

7.2 Algorithms Based on Low-Rank Approximation of the Gram Matrix. Another possibility is to replace the Gram matrix by a low-rank approximation by using greedy algorithms or randomization techniques (Smola & Schölkopf, 2000; Drineas & Mahoney, 2005). Low-rank approximations can be used, for example, to find a sparse suboptimal solution to Tikhonov regularization (Smola & Schölkopf, 2000).

7.3 Algorithms Based on Convex Formulations of the Problem. Also, when the functional to be minimized is convex and defined on a convex set, when suboptimal solutions in span \( D \) are searched for, the corresponding optimization problem may not be convex any more. Then one may consider a related convex optimization problem with sparse optimal solutions, for which efficient convex optimization algorithms can be exploited. In linear regression, for example, adding an upper bound on the \( l_1 \)-norm of the coefficients instead of their \( l_2 \)-norm (or an \( l_1 \) penalization term instead of an \( l_2 \) one) is known to enforce the sparseness of the solution. This is the least absolute shrinkage and selection operator (LASSO) problem (Tibshirani, 1996), for

---

4The algorithms in Das and Kempe (2008), Gilbert et al. (2003), Gribonval and Vandergheynst (2006), and Tropp (2004) require the knowledge of the scalar products between the function to be approximated (in our case \( f^\alpha_z \)) and the elements of the dictionary (in our case \( G_{K\alpha} \)).
which kernel versions have been proposed (Roth, 2004). In Efron, Hastie, Johnstone, and Tibshirani (2004), an algorithm well-suited to LASSO was proposed, which shows how the degree of sparseness of its solution is controlled by varying the regularization parameter. Another computationally promising technique to solve LASSO is based on the operator-splitting approach studied in Hale, Yin, and Zhang (2008). Some limitations of LASSO have been overcome by its extension called elastic net (Zou & Hastie, 2005; De Mol, De Vito, & Rosasco, 2009), where the $l_1$ and $l_2$ penalization terms are simultaneously present. Zou and Hastie (2005), showed that the elastic net can be considered as a LASSO on an extended artificial data set, so that the algorithms from Efron et al. (2004) can be still applied.

Appendix: SLT Bounds in Terms of Rademacher’s Complexity

For a loss function $V$ and a family $F$ of functions $f : X \rightarrow \mathbb{R}$, typically SLT bounds have either the two-sided form

$$ P \left\{ \sup_{f \in F} |E_V(f) - E_{z,V}(f)| \leq \delta(m, q, \tau) \right\} \geq 1 - \tau $$

or the one-sided form

$$ P \left\{ \sup_{f \in F} (E_V(f) - E_{z,V}(f)) \leq \delta(m, q, \tau) \right\} \geq 1 - \tau, $$

where $P$ denotes the a priori probability with respect to the random draw of the i.i.d. data sample $z$ of size $m$, $\tau \in (0, 1)$, $\delta(m, q, \tau)$ is a suitable nonnegative function, and the parameter $q$ measures the complexity of the family $F$ (e.g., its Rademacher’s complexity, its covering numbers, its VC dimension). We are interested in SLT bounds formulated in terms of the Rademacher’s complexity of $F$, defined, for a probability measure $\rho_X$ generating the i.i.d. input data sample $x = (x_1, \ldots, x_m) \in X^m$ and $m$ independent, uniformly distributed, and $\{\pm 1\}$-valued random variables $\sigma = \{\sigma_1, \ldots, \sigma_m\}$, as

$$ R_m(F) = E_x E_{\sigma} \left\{ \sup_{f \in F} \left| \frac{1}{m} \sum_{i=1}^{m} \sigma_i f(x_i) \right| \middle| x_1, \ldots, x_m \right\} $$

(A.1)

(see Bartlett & Mendelson, 2002; Shawe-Taylor & Cristianini, 2004; a slightly different definition is given in Mendelson, 2003). For a given $x \in X^m$, the empirical Rademacher’s complexity of $F$ is defined as

$$ \hat{R}_m(F) = E_{\sigma} \left\{ \sup_{f \in F} \left| \frac{1}{m} \sum_{i=1}^{m} \sigma_i f(x_i) \right| \middle| x_1, \ldots, x_m \right\}. $$

(A.2)
Thus,

\[ R_m(\mathcal{F}) = \mathbb{E}_z[\hat{R}_m(\mathcal{F})]. \quad (A.3) \]

The following result is a restatement of Bartlett and Mendelson (2002, theorem 8) and Shawe-Taylor and Cristianini (2004, theorem 4.9).

**Theorem 5.** Let \( L > 0, V : \mathbb{R}^2 \to [0, L] \) a loss function, \( \mathcal{F} \) a family of functions \( f : X \to \mathbb{R} \), \( z \) an i.i.d. data sample of size \( m \), and \( \tilde{V} \circ \mathcal{F} = \{(x, y) \to V(f(x), y) - V(0, y) : f \in \mathcal{F}\} \). Then for every positive integer \( m \) and every \( \tau \in (0, 1) \),

\[
P \left\{ \sup_{f \in \mathcal{F}} \frac{\mathcal{E}_V(f) - \mathcal{E}_z V(f)}{L} \leq \frac{1}{L} R_m(\tilde{V} \circ \mathcal{F}) + \sqrt{\frac{\ln(2/\tau)}{2m}} \right\} \geq 1 - \tau, \quad (A.4)
\]

where \( P \) denotes the a priori probability with respect to the random draw of the data.

When \( V \) is the square-loss function and \( \mathcal{F} \) is a bounded subset of a RKHS \( \mathcal{H}_K(X) \), Theorem 5 gives the following corollary, which is a restatement of Shawe-Taylor and Cristianini (2004, theorem 7.39).\(^5\) Its proof exploits several structural results on Rademacher’s complexity and its empirical version (see Bartlett & Mendelson, 2002, theorem 12, or theorem 4.15) and a probabilistic upper bound on the empirical Rademacher’s complexity of the ball \( B_R(\| \cdot \|_K) = \{ f \in \mathcal{H}_K(X) : \| f \|_K \leq R \} \) of radius \( R \) in the RKHS \( \mathcal{H}_K(X) \) (see Bartlett & Mendelson, 2002, lemma 22, or Shawe-Taylor & Cristianini, 2004, theorem 4.12):

**Theorem 6.** Let \( X \) be a nonempty set, \( K : X \times X \to \mathbb{R} \) a psd kernel, \( s_K = \sup_{x \in X} \sqrt{K(x, x)} < +\infty \), \( 0 < R_1 \leq R_2 \), \( z \) an i.i.d. data sample of size \( m \), and assume that the support of the marginal probability measure \( \rho_Y \) associated with the random variable \( Y \) is a subset of \([-s_K R_2, s_K R_2]\). Then for every positive integer \( m \) and every \( \tau \in (0, 1) \),

\[
P \left\{ \sup_{f \in B_{R_2}(\| \cdot \|_K)} (\mathcal{E}(f) - \mathcal{E}_z(f)) \leq \frac{16 s_K R_2}{m} \left( R_1 \sqrt{\text{tr}(K[\mathbf{x}]) + \| \mathbf{y} \|_2} \right) + 12 (s_K R_2)^2 \sqrt{\frac{\ln(2/\tau)}{2m}} \right\} \geq 1 - \tau,
\]

\(^5\)In the statement from Shawe-Taylor and Cristianini (2004, theorem 7.39), there is only one constant \( R > 0 \) instead of the two positive constants \( R_1, R_2 \), that is, one takes \( R_1 = R_2 = R > 0 \) in theorem 6. Inspection of the proof (see Shawe-Taylor & Cristianini, 2004, pp. 231–232); shows that one can restate the theorem in terms of two different constants \( R_1 \) and \( R_2 \), provided that \( 0 < R_1 \leq R_2 \).
where \( P \) denotes the a priori probability with respect to the random draw of the data.

The bound in theorem 6 contains a data-dependent term \( (R_1 \sqrt{\text{tr}(K[x])} + \|y\|_2) \), which disappears in the following slight variation of the estimate:

**Theorem 7.** Under the same assumptions of theorem 6, if \( X \) is compact and \( K \) is continuous, then

\[
\Pr \left\{ \sup_{f \in B_{R_1}(\|\cdot\|_K)} (\mathcal{E}(f) - \mathbb{E}_z(f)) \leq \frac{16s_K^2 R_2}{\sqrt{m}} (R_1 + R_2) + 4(s_K R_2)^2 \sqrt{\frac{\ln(2/\tau)}{2m}} \right\} \geq 1 - \tau.
\]

**Proof.** In the proof of Shawe-Taylor and Cristianini (2004, theorem 7.39), the probabilistic bound \( \sup_{f \in \mathcal{F}} \frac{\mathbb{E}_V(f) - \mathbb{E}_z(f)}{L} \leq \frac{1}{L} \hat{R}_m(\mathcal{V} \circ \mathcal{F}) + 3\sqrt{\frac{\ln(2/\tau)}{2m}} \) from equation A.4 is exploited. Using instead the bound \( \sup_{f \in \mathcal{F}} \frac{\mathbb{E}_V(f) - \mathbb{E}_z(f)}{L} \leq \frac{1}{L} R_m(\mathcal{V} \circ \mathcal{F}) + \sqrt{\frac{\ln(2/\tau)}{2m}} \), again from equation A.4, together with equation A.2, and then proceeding as in the proof of Shawe-Taylor and Cristianini (2004, theorem 7.39), we replace the term \( R_1^{-1/2} \sqrt{\text{tr}(K[x])} \) by \( \sqrt{\text{tr}(T_K)} \) (as in Bartlett & Mendelson, 2002, section 4.3), where \( T_K \) is the integral operator defined as \( T_K(f) = \int_X K(x, t) f(t) d\rho_X(t) \). By the corollary to the Mercer theorem stated in Cucker and Smale (2001), we get \( \text{tr}(T_K) = \int_X K(t, t) d\rho_X(t) \leq s_K^2 \). Similarly, by equation A.2 and the assumption on the support of \( \rho_Y \), the term \( \|y\|_2^2/m \) is replaced by \( s_K R_2/\sqrt{m} \).

**Acknowledgments**

We are grateful to an anonymous reviewer for comments and suggestions that allowed us to improve the original manuscript. We were partially supported by the grant Progetti di Ricerca di Ateneo 2008, from the University of Genova for the project Solution of Functional Optimization Problems by Nonlinear Approximators and Learning from Data.

**References**


Received May 16, 2008; accepted June 23, 2009.