

# Infinite- $\sigma$ Limits For Tikhonov Regularization

**Ross A. Lippert**

*Department of Mathematics  
Massachusetts Institute of Technology  
77 Massachusetts Avenue  
Cambridge, MA 02139-4307, USA*

LIPPERT@MATH.MIT.EDU

**Ryan M. Rifkin**

*Honda Research Institute USA, Inc.  
145 Tremont Street  
Boston, MA 02111, USA*

RRIFKIN@HONDA-RI.COM

**Editor:** Gabor Lugosi

## Abstract

We consider the problem of Tikhonov regularization with a general convex loss function: this formalism includes support vector machines and regularized least squares. For a family of kernels that includes the Gaussian, parameterized by a “bandwidth” parameter  $\sigma$ , we characterize the limiting solution as  $\sigma \rightarrow \infty$ . In particular, we show that if we set the regularization parameter  $\lambda = \tilde{\lambda}\sigma^{-2p}$ , the regularization term of the Tikhonov problem tends to an indicator function on polynomials of degree  $\lfloor p \rfloor$  (with residual regularization in the case where  $p \in \mathbb{Z}$ ). The proof rests on two key ideas: *epi-convergence*, a notion of functional convergence under which limits of minimizers converge to minimizers of limits, and a *value-based formulation of learning*, where we work with regularization on the function output values ( $y$ ) as opposed to the function expansion coefficients in the RKHS. Our result generalizes and unifies previous results in this area.

**Keywords:** Tikhonov regularization, Gaussian kernel, theory, kernel machines

## 1. Introduction

Given a data set  $(x_1, \hat{y}_1), \dots, (x_n, \hat{y}_n) \in \mathbb{R}^d \times \mathbb{R}$ , the supervised learning task is to construct a function  $f(x)$  that, given a new point,  $x$ , will predict the associated  $y$  value. A number of methods for this problem have been studied. One popular family of techniques is Tikhonov regularization in a Reproducing Kernel Hilbert Space (RKHS) (Evgeniou et al., 2000):

$$\inf_{f \in \mathcal{H}} \left\{ n\lambda \|f\|_{\kappa}^2 + \sum_{i=1}^n v(f(x_i), \hat{y}_i) \right\}.$$

Here,  $v : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$  is a *loss function* indicating the price we pay when we see  $x_i$ , predict  $f(x_i)$ , and the true value is  $\hat{y}_i$ . The squared norm,  $\|f\|_{\kappa}^2$ , in the RKHS  $\mathcal{H}$  involves the kernel function  $\kappa$  (Aronszajn, 1950). The regularization constant,  $\lambda > 0$ , controls the trade-off between fitting the training set accurately (minimizing the penalties) and forcing  $f$  to be smooth in  $\mathcal{H}$ . The Representer Theorem (Wahba, 1990; Girosi et al., 1995; Schölkopf et al., 2001) guarantees that the solution to

the Tikhonov regularization can be written in the form

$$f(x) = \sum_{i=1}^n c_i \kappa(x_i, x).$$

In practice, solving a Tikhonov regularization problem is equivalent to finding the expansion coefficients  $c_i$ .

One popular choice for  $\kappa$  is the *Gaussian* kernel  $\kappa_\sigma(x, x') = e^{-\frac{\|x-x'\|^2}{2\sigma^2}}$ , where  $\sigma$  is the bandwidth of the Gaussian. Common choices for  $v$  include the *square* loss,  $v(y, \hat{y}) = (y - \hat{y})^2$ , and the *hinge* loss,  $v(y, \hat{y}) = \max\{0, 1 - y\hat{y}\}$ , which lead to regularized least squares and support vector machines, respectively.

Our work was originally motivated by the empirical observation that on a range of tasks, regularized least squares achieved very good performance with very large  $\sigma$ . (For example, we could often choose  $\sigma$  so large that every kernel product between pairs of training points was between .99999 and 1.) To get good results with large  $\sigma$ , it was necessary to make  $\lambda$  small. We decided to study this relationship.

Regularized least squares (RLS) is an especially simple Tikhonov regularization algorithm: “training” RLS simply involves solving a system of linear equations. In particular, defining the matrix  $K$  via  $K_{ij} = \kappa(x_i, x_j)$ , the RLS expansion coefficients  $c$  are given by  $(K + n\lambda I)c = \hat{y}$ , or  $c = (K + n\lambda I)^{-1}\hat{y}$ . Given a test point  $x_0$ , we define the  $n$ -vector  $k$  via  $k_i = \kappa(x_0, x_i)$ , and we have, for RLS with a fixed bandwidth,

$$f(x_0) = k^t (K + n\lambda I)^{-1} \hat{y}.$$

In Lippert and Rifkin (2006), we studied the limit of this expression as  $\sigma \rightarrow \infty$ , showing that if we set  $\lambda = \tilde{\lambda}\sigma^{-2p-1}$  for  $p$  a positive integer, the infinite- $\sigma$  limit converges (pointwise) to the degree  $p$  polynomial with minimal empirical risk on the training set. The asymptotic predictions are equivalent to those we would get if we simply fit an (unregularized) degree  $p$  polynomial to our training data.

In Keerthi and Lin (2003), a similar phenomenon was also noticed for support vector machines (SVM) with Gaussian kernels, where it was observed that the SVM function could be made to converge (in the infinite- $\sigma$  limit) to a linear function which minimized the hinge loss plus a residual regularization (discussed further below). In that work, only a linear result was obtained; no results were given for general polynomial approximation limits.

In the current work, we unify and generalize these results, showing that the occurrence of these polynomial approximation limits is a general phenomenon, which holds across all convex loss functions and a wide variety of kernels taking the form  $\kappa_\sigma(x, x') = \kappa(x/\sigma, x'/\sigma)$ . Our main result is that for a convex loss function and a valid kernel, if we take  $\sigma \rightarrow \infty$  and  $\lambda = \tilde{\lambda}\sigma^{-2p}$ , the regularization term of the Tikhonov problem tends to an indicator function on polynomials of degree  $\lfloor p \rfloor$ . In the case where  $p \in \mathbb{Z}$ , there is residual regularization on the degree- $p$  coefficients of the limiting polynomial.

Our proof relies on two key ideas. The first is the notion of *epi-convergence*, a functional convergence under which limits of minimizers converge to minimizers of limits. This notion allows us to characterize the limiting Tikhonov regularization problem in a mathematically precise way. The second notion is a *value-based formulation of learning*. The idea is that instead of working with the expansion coefficients in the RKHS ( $c_i$ ), we can write the regularization problem directly in terms of the predicted values ( $y_i$ ). This allows us to avoid combining and canceling terms whose limits are individually undefined.

## 2. Notation

In this section, we describe the notation we use throughout the paper. Some of our choices are non-standard, and we try to indicate these.

### 2.1 Data Sets and Regularization

We refer to a general  $d$  dimensional vector with the symbol  $x$  (plus superscripts or subscripts). We assume a fixed set of  $n$  training points  $x_i \in \mathbb{R}^d$  ( $1 \leq i \leq n$ ) and refer to the totality of these data points by  $X$ , such that for any  $f : \mathbb{R}^d \rightarrow \mathbb{R}$ ,  $f(X) = (f(x_1) \cdots f(x_n))^t$  is the vector of values over the points and for any  $g : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ ,  $g(X, X)$  is the matrix of values over pairs of points, i.e.  $[g(X, X)]_{ij} = g(x_i, x_j)$ . We let  $x_0$  represent an arbitrary “test point” not in the training set. We assume we are given “true”  $\hat{y}$  values at the training points:  $\hat{y}_i \in \mathbb{R}, 1 \leq i \leq n$ . While it is more common to use  $\hat{y}$  to refer to the “predicted” output values and  $y$  to refer to the “true” output values, we find this choice much more notationally convenient, because our value-based formulation of learning (3) requires us to work with the predicted values very frequently.

Tikhonov regularization is given by

$$\inf_{f \in \mathcal{H}} \left\{ n\lambda \|f\|_{\mathbb{K}}^2 + \sum_{i=1}^n v(f(x_i), \hat{y}_i) \right\}. \quad (1)$$

Tikhonov regularization can be used for both classification and regression tasks, but we refer to the function  $f$  as the *regularized solution* in all cases. We call the left-hand portion the *regularization term*, and the right-hand portion the *loss term*. We assume a loss function  $v(y, \hat{y})$  that is convex in its first argument and minimized at  $y = \hat{y}$  (thereby ruling out, for example, the 0/1 “misclassification rate”). We call such a loss function *valid*. Aside from convexity, we will be unconcerned with the form of the loss function and often take the loss term in the optimization in (1) to be some convex function  $V : \mathbb{R}^n \rightarrow \mathbb{R}$  which is minimized by the vector  $\hat{y}$  of  $\hat{y}_i$ 's.

To avoid confusion, when subscripting over  $d$  dimensional indices, we use letters from the beginning of the alphabet ( $a, b, c, \dots$ ), while using letters from the middle ( $i, j, \dots$ ) for  $n$  dimensional indices.

When referring to optimization problems, we use an over-dot (e.g.  $\dot{y}$ ) to denote optimizing quantities. We are not time-differentiating anything in this work, so this should not cause confusion.

### 2.2 Polynomials

By a *multi-index* we refer to  $I \in \mathbb{Z}^d$  such that  $I_a \geq 0$ . Given  $x \in \mathbb{R}^d$  we write  $x^I = \prod_{a=1}^d x_a^{I_a}$ . We also write  $X^I$  to denote  $(x_1^I \cdots x_n^I)^t$ . The *degree* of a multi-index is  $|I| = \sum_{a=1}^d I_a$ . We use the “choose” notation

$$\binom{|I|}{I} = \frac{|I|!}{\prod_{a=1}^d I_a!}.$$

Let  $\{I_i\}_{i=0}^\infty$  be an ordering of multi-indices which is non-decreasing by degree (in particular,  $I_0 = (0 \cdots 0)$ ). We consider this fixed for the remainder of the work. Define  $d_c = |\{I : |I| \leq c\}|$  and note that  $d_c = \binom{d+c+1}{c}$ . Put differently,  $\{I : |I| = c\} = \{I_i : d_{c-1} \leq i < d_c\}$ .

Given a data set, while the monomials  $x^I$  are linearly independent as functions, no more than  $n$  of the  $X^I$  can be linearly independent. We say that a data set is *generic* if the  $X^I$ , for  $i < n$ , are

linearly independent. This is equivalent to requiring that the data not reside on the zero-set of a low degree system polynomials. This is not an unreasonable assumption for data which is presumed to have been generated by distributions supported on  $\mathbb{R}^n$  or some sphere in  $\mathbb{R}^n$ . Throughout this paper, we assume that our data is generic. It is possible to carry out the subsequent derivations without it, but the modifications which result are tedious. In particular, parts of Theorem 12, which treat the first  $n$  monomials  $X^{I_i}$  as linearly independent (e.g. assuming  $v_\alpha(X)$  is non-singular in the proof), would need to be replaced with analogous statements about the first  $n$  monomials  $X^{I_j}$  which are linearly independent, and various power-series expansion coefficients would have to be adjusted accordingly. Additionally, our main result requires not only genericity of the data, but also that  $n > d_p$  where  $p$  is the degree of the asymptotic regularized solution.

### 2.3 Kernels

It is convenient to use infinite matrices and vectors to express certain infinite series. Where used, the convergence of the underlying series implies the convergence of any infinite sums that arise from the matrix products we form, and we will not attempt to define any inverses of non-diagonal infinite matrices. This is merely a notational device to avoid excessive explicit indexing and summing in the formulas ahead. Additionally, since many of our vectors come from power series expansions, we adopt the convention of indexing such vectors and matrices starting from 0.

A Reproducing Kernel Hilbert Space (RKHS) is characterized by a kernel function  $\kappa$ . If  $\kappa$  has a power series expansion, we may write

$$\begin{aligned}\kappa(x, x') &= \sum_{i, j \geq 0} M_{ij} x^i x'^j \\ &= v(x) M v(x')^t\end{aligned}$$

where  $M \in \mathbb{R}^{\infty \times \infty}$  is an infinite matrix and  $v(x) = (1 \quad x^{I_1} \quad x^{I_2} \quad \dots) \in \mathbb{R}^{1 \times \infty}$  is an infinite row-vector valued function of  $x$ . We emphasize that  $M$  is an infinite matrix induced by the kernel function  $\kappa$  and the ordering of multi-indices; it has nothing to do with our data set.

We say that a kernel is *valid* if every finite upper-left submatrix of  $M$  is symmetric and positive definite; in this case, we also say that the infinite matrix  $M$  is symmetric positive definite. This condition is the one we use in our main proof; however, it can be difficult to check. It is independent of the Mercer property (which states that the kernel matrix  $\kappa(X, X)$  for a set  $X$  is positive semidefinite), since  $\kappa(x, x') = \frac{1}{1-xx'}$  is valid but not Mercer, and  $\exp(-(x^3 - x'^3)^2)$  is Mercer but not valid. This notion is, basically, that the feature space of  $\kappa$  can approximate any polynomial function near the origin to arbitrary accuracy. We are not aware of any mention of this property in the literature. The following lemma gives a stronger condition that implies validity.

**Lemma 1** *If  $\kappa(x, x') = \sum_{c \geq 0} (x \cdot x')^c g_c(x) g_c(x')$  for some analytic functions  $g_c(x)$  such that  $g_c(0) \neq 0$ , then  $\kappa$  is a valid kernel.*

**Proof** By the multinomial theorem,

$$(x \cdot x')^c = \left( \sum_{i=1}^d x_i x'_i \right)^c = \sum_{|I|=c} \binom{|I|}{I} x^I x'^I.$$

Let  $g_c(x) = \sum_I G_{cI} x^I$ , and note  $g_c(0) = G_{c0}$ , thus

$$\begin{aligned} \kappa(x, x') &= \sum_{I, J, c \geq 0} \sum_{|E|=c} \binom{|E|}{E} x^{I+E} G_{cI} G_{cJ} x'^{J+E} \\ &= \sum_{I, J, E} \binom{|I_k|}{I_k} x^{I+E} G_{|E|I} G_{|E|J} x'^{J+E} \\ &= \sum_{I, J \geq E} x^I G_{|E|(I-E)} \binom{|E|}{E} G_{|E|(J-E)} x'^J \\ &= v(x) L L^t v(x') \end{aligned}$$

where  $L_{ij} = \binom{|I_j|}{I_j}^{\frac{1}{2}} G_{|I_j|(I_i-I_j)}$  when  $I_i \geq I_j$  and 0 otherwise. In other words,  $L$  is an infinite lower triangular matrix with non-vanishing diagonal elements (since  $G_{c0} \neq 0$  for all  $c$ ).

Since  $M = L L^t$ , the upper-left submatrices of  $L$  are the Cholesky factors of the corresponding upper-left submatrix of  $M$ , and thus  $M$  is positive definite.  $\blacksquare$

We note that  $\kappa(x, x') = \exp(-\frac{1}{2}\|x - x'\|^2)$  can be written in the form of Lemma 1:

$$\begin{aligned} \kappa(x, x') &= \exp\left(-\frac{1}{2}\|x\|^2\right) \exp(x \cdot x') \exp\left(-\frac{1}{2}\|x'\|^2\right) \\ &= \sum_{c=0}^{\infty} \frac{(x \cdot x')^c}{c!} \exp\left(-\frac{1}{2}\|x\|^2\right) \exp\left(-\frac{1}{2}\|x'\|^2\right) \\ &= \sum_{c=0}^{\infty} (x \cdot x')^c g_c(x) g_c(x'), \end{aligned}$$

where  $g_c(x) = \frac{1}{\sqrt{c!}} \exp(-\frac{1}{2}\|x\|^2)$ .

We will consider kernel functions  $\kappa_\sigma$  which are parametrized by a bandwidth parameter  $\sigma$  (or  $s = \frac{1}{\sigma}$ ). We will occasionally use  $K_\sigma$  to refer to the matrix whose  $i, j$ th entry  $\kappa_\sigma(x_i, x_j)$ , for  $1 \leq i \leq n, 1 \leq j \leq n$ . We will also use  $k_\sigma$  to denote the  $n$ -vector whose  $i$ th entry is  $\kappa_\sigma(x_i, x_0)$  — the kernel product between the  $i$ th training point  $x_i$  and the test point  $x_0$ .

### 3. Value-Based Learning Formulation

In this section, we discuss our value-based learning formulation. Using the representer theorem and basic facts about RKHS, the standard Tikhonov regularization problem (1) can be written in terms of the expansion coefficients  $c$  and the kernel matrix  $K$ :

$$\inf_{c \in \mathbb{R}^n} \left\{ n\lambda c^t K c + \sum_{i=1}^n v([Kc]_i, \hat{y}_i) \right\}. \quad (2)$$

The predicted values on the training set are  $y = f(X) = Kc$ . If the kernel matrix  $K$  is invertible (which is the case for a Gaussian kernel and a generic data set), then  $c = K^{-1}y$ , and we can rewrite the minimization as

$$\inf_{y \in \mathbb{R}^n} \left\{ n\lambda y^t K^{-1} y + V(y) \right\}. \quad (3)$$

where  $V$  is convex ( $V(y) = \sum_{i=1}^n v_i(y_i, \hat{y}_i)$ ).

While problem 2 is explicit in the coefficients of the expansion of the regularized solution, problem 3 is explicit in the predicted values  $y_i$ . The purpose behind our choice of formulation is to avoid the unnecessary complexities which result from replacing  $\kappa$  with  $\kappa_\sigma$  and taking limits as both  $c_i$  and  $\kappa_\sigma(x, x_i)$  change separately with  $\sigma$ : note that in problem 3, *only the regularization term is varying with  $\sigma$* .

In this section, we will show how our formulation achieves this, by allowing us to state a single optimization problem which simultaneously solves the Tikhonov regularization problem on the training data and evaluates the resulting function on the test data.

**Theorem 2** Let  $y = \begin{pmatrix} y_0 \\ y_1 \end{pmatrix} \in \mathbb{R}^{m+n}$  be a block vector with  $y_0 \in \mathbb{R}^m, y_1 \in \mathbb{R}^n$  and  $K = \begin{pmatrix} K_{00} & K_{01} \\ K_{10} & K_{11} \end{pmatrix} \in \mathbb{R}^{(m+n) \times (m+n)}$  be a positive definite matrix. For any  $V : \mathbb{R}^n \rightarrow \mathbb{R}$ , if  $\hat{y}$  minimizes

$$y^t K^{-1} y + V(y_1) \quad (4)$$

then  $\hat{y}_1$  minimizes

$$y_1^t K_{11}^{-1} y_1 + V(y_1) \quad (5)$$

and  $\hat{y}_0 = K_{01} K_{11}^{-1} \hat{y}_1$ .

**Proof**

$$\inf_{y_0, y_1} \{y^t K^{-1} y + V(y_1)\} = \inf_{y_1} \{ \inf_{y_0} \{y^t K^{-1} y\} + V(y_1) \}. \quad (6)$$

Let  $K^{-1} = \bar{K} = \begin{pmatrix} \bar{K}_{00} & \bar{K}_{01} \\ \bar{K}_{10} & \bar{K}_{11} \end{pmatrix}$ . Consider minimizing  $y^t K^{-1} y = y_0^t \bar{K}_{00} y_0 + 2y_0^t \bar{K}_{01} y_1 + y_1^t \bar{K}_{11} y_1$ . For fixed  $y_1$ ,  $\hat{y}_0 = -\bar{K}_{00}^{-1} \bar{K}_{01} y_1 = K_{01} K_{11}^{-1} y_1$  by (17) of Lemma 15. Thus,

$$\inf_{y_0} \{y^t K^{-1} y\} = y_1^t (\bar{K}_{11} - \bar{K}_{10} \bar{K}_{00}^{-1} \bar{K}_{01}) y_1 = y_1^t K_{11}^{-1} y_1$$

by (19) of Lemma 15. ■

We contextualize this result in terms of the Tikhonov learning problem with the following corollary.

**Corollary 3** Let  $X$  be a given set of data points  $x_1, \dots, x_n$  with  $x_0$  a test point. Let  $\kappa$  be a valid kernel function and  $V : \mathbb{R}^n \rightarrow \mathbb{R}$  be arbitrary. If  $\hat{y} = (\hat{y}_0 \quad \hat{y}_1 \quad \dots \quad \hat{y}_n)^t$  is the minimizer of

$$n\lambda y^t \kappa \left( \begin{pmatrix} x_0 \\ X \end{pmatrix}, \begin{pmatrix} x_0 \\ X \end{pmatrix} \right)^{-1} y + V(y_1, \dots, y_n)$$

then  $(\hat{y}_1 \quad \dots \quad \hat{y}_n)^t$  minimizes  $n\lambda y^t \kappa(X, X)^{-1} y + V(y)$  and  $\hat{y}_0 = \sum_{i=1}^n c_i \kappa(x_0, x_i)$ , for  $c = \kappa(X, X)^{-1} (\hat{y}_1 \quad \dots \quad \hat{y}_n)^t$ .

Thus, when solving for  $y$  instead of  $c$ , we can evaluate the function at a test point  $x_0$  by including the additional point in a larger minimization problem where the test point contributes to the regularization, but not the loss. When taking limits, we are going to work directly with the  $y_i$ , and we are going to avoid dealing with the (divergent) limits of the  $c_i$ .

#### 4. Epi-limits, Convex Functions, and Quadratic Forms

The relationship between the limit of a function and the limit of its minimizer(s) is subtle, and it is very easy to make incorrect statements. For convex functions there are substantial results on this subject, which we review; we essentially follow the development of Rockafellar and Wets (2004, chap. 7). Since the component of our objective which depends on the limiting parameter is a quadratic form, we will eventually specialize the results to quadratic forms.

**Definition 4 (epigraphs)** Given a function  $f : \mathbb{R}^n \rightarrow (-\infty, \infty]$ , its epigraph, *epi*  $f$  is the subset of  $\mathbb{R} \times \mathbb{R}^n$  given by

$$\text{epi } f = \{(z, x) : z \geq f(x)\}.$$

We call  $f$  closed, convex, or proper if those statements are true of *epi*  $f$  (proper referring to *epi*  $f$  being neither  $\emptyset$  nor  $\mathbb{R}^{n+1}$ ).

The functions we will be interested in are closed, convex, and proper. We will therefore adopt the abbreviation *ccp* for these conditions. Additionally, since we will be studying parameterized functions,  $f_s$ , for  $0 < s$  as  $s \rightarrow 0$ , we say that such a family of functions is *eventually* convex (or closed, or proper) when there exists some  $s_0 > 0$  such that  $f_s$  is convex (or closed, or proper) for all  $0 < s < s_0$ .

We review the definition of  $\liminf$  and  $\limsup$  for functions of a single variable. Given  $h : (0, \infty) \rightarrow (-\infty, \infty]$ , it is clear that the functions  $\inf_{s' \in (0, s)} \{h(s')\}$  and  $\sup_{s' \in (0, s)} \{h(s')\}$  are non-increasing and non-decreasing functions of (increasing)  $s$  respectively.

**Definition 5** For  $h : (0, \infty) \rightarrow (-\infty, \infty]$ ,

$$\begin{aligned} \liminf_{s \rightarrow 0} h(s) &= \sup_{s > 0} \left\{ \inf_{s' \in (0, s)} \{h(s')\} \right\} \\ \limsup_{s \rightarrow 0} h(s) &= \inf_{s > 0} \left\{ \sup_{s' \in (0, s)} \{h(s')\} \right\}. \end{aligned}$$

As defined, either of the limits may take the value  $\infty$ . A useful alternate characterization, which is immediate from the definition, is  $\liminf_{s \rightarrow 0} h(s) = h_0$  iff  $\forall \varepsilon > 0, \exists s_0, \forall s \in (0, s_0) : h(s) \geq h_0 - \varepsilon$ , and  $\limsup_{s \rightarrow 0} h(s) = h_0$  iff  $\forall \varepsilon > 0, \exists s_0, \forall s \in (0, s_0) : h(s) \leq h_0 + \varepsilon$ , where either inequality can be strict if  $h_0 < \infty$ .

**Definition 6 (epi-limits)** We say  $\lim_{s \rightarrow 0} f_s = f$  if for all  $x_0 \in \mathbb{R}^n$ , both the following properties hold:  
Property 1:  $\forall x : [0, \infty) \rightarrow \mathbb{R}^n$  continuous at  $x(0) = x_0$  satisfies

$$\liminf_{s \rightarrow 0} f_s(x(s)) \geq f(x_0) \tag{7}$$

Property 2:  $\exists x : [0, \infty) \rightarrow \mathbb{R}^n$  continuous at  $x(0) = x_0$  satisfying

$$\limsup_{s \rightarrow 0} f_s(x(s)) \leq f(x_0). \tag{8}$$

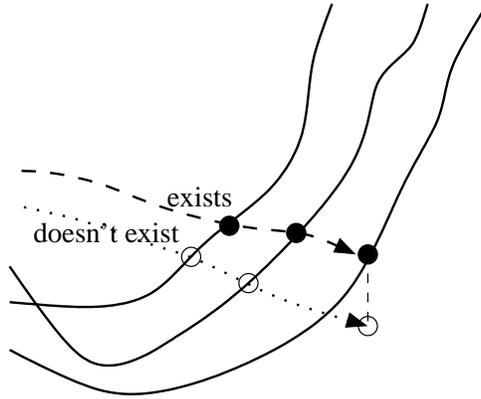


Figure 1: (property 1) of Definition 6 says that paths of points within  $\text{epi } f_s$  cannot end up below  $\text{epi } f$ , while (property 2) says that at least one such path hits every point of  $\text{epi } f$ .

This notion of functional limit is called an epigraphical limit (or epi-limit). Less formally, (property 1) is the condition that paths of the form  $(x(s), f_s(x(s)))$  are, asymptotically, inside  $\text{epi } f$ , while (property 2) asserts the existence of a path which hits the boundary of  $\text{epi } f$ , as depicted in figure 1. Considering (property 1) with the function  $x(s) = x_0$ , it is clear that the epigraphical limit minorizes the pointwise limit (assuming both exist), but the two need not coincide. An example of this distinction is given by the family of functions

$$f_s(x) = \frac{2}{s}x(x-s) + 1,$$

illustrated by Figure 2. The pointwise limit is  $f_0(0) = 1, f_0(x) = \infty$  for  $x \neq 0$ . The epi-limit is 0 at 0.

We say that a quadratic form is *finite* if  $f(x) < \infty$  for all  $x$ . (We note in passing that if a quadratic form is not finite,  $f(x) = \infty$  almost everywhere.) The pointwise and epi-limits of quadratic forms agree when the limiting quadratic form is finite, but the example in the figure is not of that sort. This behavior is typical of the applications we consider. In what follows, we take all functional limits to be epi-limits.

It is the epi-limit of functions which is appropriate for optimization theory, as the following theorem (a variation of one one from Rockafellar and Wets (2004)) shows.

**Theorem 7** *Let  $f_s : \mathbb{R}^n \rightarrow (-\infty, \infty]$  be eventually ccp, with  $\lim_{s \rightarrow 0} f_s = f$ . If  $f_s, f$  have unique minimizers  $\hat{x}(s), \hat{x}$  then*

$$\lim_{s \rightarrow 0} \hat{x}(s) = \hat{x} \quad \text{and} \quad \lim_{s \rightarrow 0} f_s(\hat{x}(s)) = \inf_x f(x).$$

**Proof** Given  $\delta > 0$ , let  $B_\delta = \{x \in \mathbb{R}^n : f(x) < f(\hat{x}) + 2\delta\}$ . Since  $\hat{x}$  is the unique minimizer of  $f$  and  $f$  is ccp,  $B_\delta$  is bounded and open, and for any open neighborhood  $U$  of  $\hat{x}$ ,  $\exists \delta > 0 : B_\delta \subset U$ . Note that  $x \in \partial B_\delta$  iff  $f(x) = f(\hat{x}) + 2\delta$ .

Let  $\hat{x} : [0, \infty) \rightarrow \mathbb{R}$  satisfy property 2 of definition 6 with  $\hat{x}(0) = \hat{x}$ . Let  $s_0 > 0$  be such that  $\forall s \in (0, s_0) : f_s(\hat{x}(s)) < f(\hat{x}) + \delta$  and  $\hat{x}(s) \in B_\delta$ . Thus,  $\forall s \in (0, s_0) : \inf_{x \in B_\delta} f_s(x) < f(\hat{x}) + \delta$ .

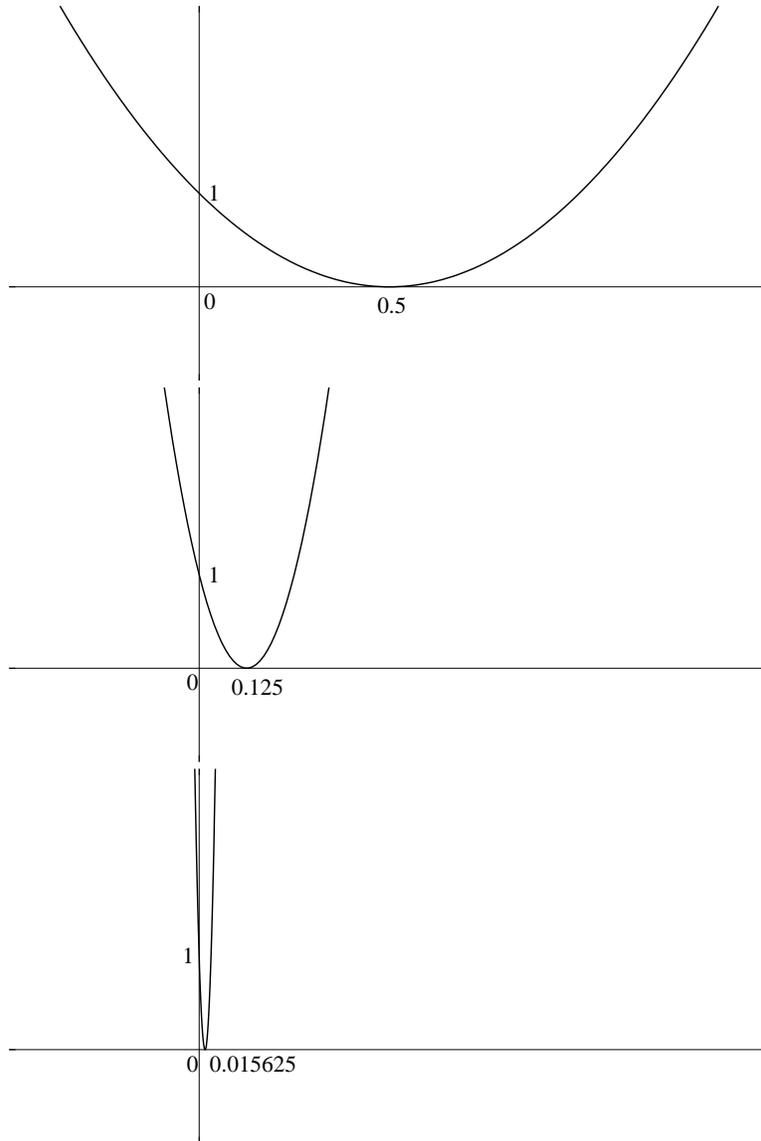


Figure 2: The function above,  $f_s(x) = \frac{2}{s}x(x-s) + 1$ , has different pointwise and epi-limits, having values 1 and 0, respectively, at  $x = 0$  and  $\infty$  for all other  $x$ .

By property 1 of definition 6,  $\forall x \in \mathbb{R}^n, \liminf_{s \rightarrow 0} f_s(x) \geq f(x)$ , in particular,  $\forall x \in \partial B_\delta, \exists s_1 \in (0, s_0), \forall s \in (0, s_1) : f_s(x) \geq f(x) - \delta = f(\dot{x}) + \delta$ . Since  $\partial B_\delta$  is compact, we can choose  $s_1 \in (0, s_0), \forall x \in \partial B_\delta, s \in (0, s_1) : f_s(x) \geq f(\dot{x}) + \delta$ .

Thus  $\forall x \in \partial B_\delta, s < s_1 : f_s(x) \geq f(\dot{x}) + \delta > \inf_{x \in B_\delta} f_s(x)$ , and therefore  $\dot{x}(s) \in B_\delta$  by the convexity of  $f_s$ .

Summarizing,  $\forall \delta > 0, \exists s_1 > 0, \forall s \in (0, s_1) : \dot{x}(s) \in B_\delta$ . Hence  $\dot{x}(s) \rightarrow \dot{x}$  and we have the first limit.

The second limit is a consequence of the first ( $\lim_{s \rightarrow 0} \dot{x}(s) = \dot{x}$ ) and definition 6. In particular,  $\limsup_{s \rightarrow 0} f_s(\dot{x}(s)) \leq f(\dot{x})$  and  $f(\dot{x}) \leq \liminf_{s \rightarrow 0} f_s(\dot{x}(s))$ . Since  $\forall s : f_s(\dot{x}(s)) \leq f_s(\hat{x}(s))$ , we have  $\limsup_{s \rightarrow 0} f_s(\hat{x}(s)) \leq f(\dot{x})$  and hence  $f(\dot{x}) \leq \liminf_{s \rightarrow 0} f_s(\hat{x}(s)) \leq \limsup_{s \rightarrow 0} f_s(\hat{x}(s)) \leq f(\dot{x})$ . ■

We now apply this theorem to characterize limits of quadratic forms (which are becoming infinite in the limit). The following lemma is elementary.

**Lemma 8** *Let  $A(s)$  be a continuous matrix-valued function. If  $A(0)$  is non-singular, then  $A(s)^{-1}$  exists for a neighborhood of  $s = 0$ .*

**Lemma 9** *Let  $Z(s) \in \mathbb{R}^{n \times n}$  be a continuous matrix valued function defined for  $s \geq 0$  such that  $Z(0) = 0$  and  $Z(s)$  is non-singular for  $s > 0$ .*

*Let  $M(s) = \begin{pmatrix} A(s) & B(s)^t \\ B(s) & C(s) \end{pmatrix} \in \mathbb{R}^{(m+n) \times (m+n)}$  be a continuous symmetric matrix valued function of  $s$  such that  $M(s)$  is positive semi-definite and  $C(s)$  is positive definite for  $s \geq 0$ . If*

$$f_s(x, y) = \begin{pmatrix} x \\ Z(s)^{-1}y \end{pmatrix}^t \begin{pmatrix} A(s) & B(s)^t \\ B(s) & C(s) \end{pmatrix} \begin{pmatrix} x \\ Z(s)^{-1}y \end{pmatrix}$$

then  $\lim_{s \rightarrow 0} f_s = f$ , where

$$f(x, y) = \begin{cases} \infty & y \neq 0 \\ x^t(A(0) - B(0)^t C(0)^{-1} B(0))x & y = 0 \end{cases}.$$

**Proof** Completing the square,

$$f_s(x, y) = \|x\|_{\tilde{A}(s)}^2 + \|Z(s)^{-1}y + C(s)^{-1}B(s)x\|_{C(s)}^2$$

where  $\|v\|_W^2 = v^t W v$  and  $\tilde{A}(s) = A(s) - B(s)^t C(s)^{-1} B(s)$ . Note that  $\tilde{A}(s)$  is positive semi-definite and continuous at  $s = 0$ .

Let  $b, c, s_0 > 0$  be chosen such that  $\forall s < s_0$

$$b\|\cdot\| > \|B(s)\cdot\|, \quad \|\cdot\|_{C(s)} > c\|\cdot\|$$

(Such quantities arise from the the singular values of the matrices involved, which are continuous in  $s$ ). Let  $z(s) = \|Z(s)\|$  (matrix 2-norm). Note:  $z$  is continuous with  $z(0) = 0$ .

Let  $x(s), y(s)$  be continuous at  $s = 0$ . If  $y(0) \neq 0$ , then for  $s < s_0$

$$\begin{aligned} \sqrt{f_s(x(s), y(s))} &\geq \|Z(s)^{-1}y(s) + C(s)^{-1}B(s)x(s)\|_{C(s)} \\ &\geq \|Z(s)^{-1}y(s)\|_{C(s)} - \|C(s)^{-1}B(s)x(s)\|_{C(s)}, \end{aligned}$$

by the triangle inequality

$$\begin{aligned}\sqrt{f_s(x(s), y(s))} &\geq \|Z(s)^{-1}y(s)\|_{C(s)} - \|B(s)x(s)\|_{C(s)^{-1}} \\ &> c \frac{\|y(s)\|}{z(s)} - \frac{b}{c} \|x(s)\| \\ &= c \left( \frac{\|y(s)\|}{z(s)} - \frac{b}{c^2} \|x(s)\| \right).\end{aligned}$$

By continuity,  $\exists s_1 \in (0, s_0)$  such that  $\forall s \in (0, s_1)$ ,

$$\|x(s)\| < \frac{3}{2} \|x(0)\|, \quad \|y(s)\| > \frac{1}{2} \|y(0)\|, \quad \frac{b}{c^2} \|x(0)\| < \frac{\|y(0)\|}{6z(s)}.$$

Thus, for all  $s < s_1$  :  $\sqrt{f_s(x(s), y(s))} > c \frac{\|y(0)\|}{4z(s)}$ , and hence  $\liminf_{s \rightarrow 0} f_s(x(s), y(s)) = \infty$ , which implies property 1 of definition 6 (and property 2, since  $\liminf \leq \limsup$ ). Otherwise ( $y(0) = 0$ ),  $f_s(x(s), y(s)) \geq \|x(s)\|_{\tilde{A}(s)}^2$  and thus

$$\begin{aligned}\lim_{s \rightarrow 0} \|x(s)\|_{\tilde{A}(s)}^2 &\leq \liminf_{s \rightarrow 0} f(x(s), y(s)) \\ \|x(0)\|_{\tilde{A}(0)}^2 &\leq \liminf_{s \rightarrow 0} f(x(s), y(s))\end{aligned}$$

(property 1).  $f_s(x(s), y(s)) = \|x(s)\|_{\tilde{A}(s)}^2$  when  $y(s) = -Z(s)C(s)^{-1}B(s)x(s)$  (which is continuous and vanishing at  $s = 0$ ), and thus

$$\limsup_{s \rightarrow 0} f(x(s), y(s)) = \lim_{s \rightarrow 0} \|x(s)\|_{\tilde{A}(s)}^2 = \|x(0)\|_{\tilde{A}(0)}^2$$

(property 2). ■

The following application of the lemma allows us to deal with matrices which will be of specific interest to us.

**Corollary 10** Let  $Z_1(s) \in \mathbb{R}^{l \times l}$  and  $Z_2(s) \in \mathbb{R}^{n \times n}$  be continuous matrix valued functions defined for  $s \geq 0$  such that  $Z_i(0) = 0$  and  $Z_i(s)$  is non-singular for  $s > 0$ .

Let  $M(s) = \begin{pmatrix} A(s) & B(s)^t & C(s)^t \\ B(s) & D(s) & E(s)^t \\ C(s) & E(s) & F(s) \end{pmatrix} \in \mathbb{R}^{(l+m+n) \times (l+m+n)}$  be a continuous symmetric matrix valued function of  $s$  such that  $M(s)$  is positive semi-definite and  $F(s)$  is positive definite for  $s \geq 0$ . If

$$f_s(q_a, q_b, q_c) = \begin{pmatrix} Z_1(s)q_a \\ q_b \\ Z_2(s)^{-1}q_c \end{pmatrix}^t \begin{pmatrix} A(s) & B(s)^t & C(s)^t \\ B(s) & D(s) & E(s)^t \\ C(s) & E(s) & F(s) \end{pmatrix} \begin{pmatrix} Z_1(s)q_a \\ q_b \\ Z_2(s)^{-1}q_c \end{pmatrix}$$

then  $\lim_{s \rightarrow 0} f_s = f$ , where

$$f(q_a, q_b, q_c) = \begin{cases} \infty & q_c \neq 0 \\ q_b^t (D(0) - E(0)^t F(0)^{-1} E(0)) q_b & q_c = 0 \end{cases}.$$

**Proof** We apply Lemma 9 to the quadratic form given by

$$\begin{pmatrix} q_a \\ q_b \\ Z_2^{-1}q_c \end{pmatrix}^t \begin{pmatrix} (Z_1^t A Z_1 & Z_1^t B^t) \\ (B Z_1 & D) \\ (C Z_1 & E) \end{pmatrix} \begin{pmatrix} (Z_1^t C^t) \\ E^t \\ F \end{pmatrix} \begin{pmatrix} q_a \\ q_b \\ Z_2^{-1}q_c \end{pmatrix}$$

( $s$  dependence suppressed). ■

We will have occasion to apply Corollary 10 when some of  $q_a, q_b$  and  $q_c$  are empty. In all cases, the appropriate result can be re-derived under the convention that a quadratic form over a 0 variables is identically 0.<sup>1</sup>

### 5. Kernel Expansions and Regularization Limits

In this section, we present our key result, characterizing the asymptotic behavior of the regularization term of Tikhonov regularization. We define a family of quadratic forms on the polynomials in  $x$ ; these forms will turn out to be the limits of the quadratic Tikhonov regularizer.

**Definition 11** Let  $\kappa(x, x') = \sum_{i,j \geq 0} M_{ij} x^i x'^j$ , with  $M$  symmetric, positive definite. For any  $p > 0$ , define  $R_p^K : f \rightarrow [0, \infty]$  by

$$R_p^K(f) = \begin{cases} 0 & f(x) = \sum_{0 \leq i \leq d_{\lfloor p \rfloor}} q_i x^i, \text{ if } p \notin \mathbb{Z} \\ \infty & \text{else} \end{cases},$$

$$R_p^K(f) = \begin{cases} \begin{pmatrix} q_{d_{p-1}+1} \\ \vdots \\ q_{d_p} \end{pmatrix}^t C \begin{pmatrix} q_{d_{p-1}+1} \\ \vdots \\ q_{d_p} \end{pmatrix} & f(x) = \sum_{0 \leq i \leq d_p} q_i x^i, \text{ if } p \in \mathbb{Z} \\ \infty & \text{else} \end{cases}$$

where, for  $p \in \mathbb{Z}$ ,  $C = (M_{bb} - M_{ba} M_{aa}^{-1} M_{ab})^{-1}$  where  $M_{aa}$  and  $\begin{pmatrix} M_{aa} & M_{ab} \\ M_{ba} & M_{bb} \end{pmatrix}$  are the  $d_{p-1} \times d_{p-1}$  and  $d_p \times d_p$  upper-left submatrices of  $K$ .

The  $q_i$  in the conditions for  $f$  above are arbitrary, and hence the conditions are both equivalent to  $f \in \text{span}\{x^I : |I| \leq p\}$ . We have written the  $q_i$  explicitly merely to define the value  $R_p^K$  when  $p \in \mathbb{Z}$ .

Define  $v(X) = (1 \quad X^{I_1} \quad X^{I_2} \quad \dots) \in \mathbb{R}^{n \times \infty}$ . Let  $v(X) = (v_\alpha(X) \quad v_\beta(X))$  be a block decomposition into an  $n \times n$  block (a *Vandermonde* matrix on the data) and an  $n \times \infty$  block. Because our data set is generic,  $v_\alpha(X)$  is non-singular, and the interpolating polynomial through the points  $(x_i, y_i)$  over the monomials  $\{x^i : i < n\}$  is given by  $f(x) = v_\alpha(x) v_\alpha(X)^{-1} y$ .

We now state and prove our key result, showing the convergence of the regularization term of Tikhonov regularization to  $R_p^K$ .

**Theorem 12** Let  $X$  be generic and  $\kappa(x, x') = \sum_{i,j \geq 0} M_{ij} x^i x'^j$  be a valid kernel. Let  $p \in [0, |I_{n-1}|)$ . Let  $f_s(y) = s^{2p} y^t \kappa(sX, sX)^{-1} y$ . Then

$$\lim_{s \rightarrow 0} f_s = f,$$

where  $f(y) = R_p^K(q)$ , and  $q(x) = v_\alpha(x) \tilde{q} = \sum_{0 \leq i < n} \tilde{q}_i x^i$  and  $\tilde{q} = v_\alpha(X)^{-1} y$ .

---

1. This is not a definition. We are merely stating (without proof) that if we were to go through the proofs omitting some of  $q_a, q_b$ , and  $q_c$ , we would obtain the same result.

**Proof** Recalling that  $v_\alpha(X)$  is non-singular by genericity, define  $\chi = v_\alpha(X)^{-1}v_\beta(X)$ . Let  $\Sigma(s)$  be the infinite diagonal matrix valued function of  $s$  whose  $i^{\text{th}}$  diagonal element is  $s^{|I_i|}$ . We define a block decomposition  $\Sigma(s) = \begin{pmatrix} \Sigma_\alpha(s) & 0 \\ 0 & \Sigma_\beta(s) \end{pmatrix}$  where  $\Sigma_\alpha(s)$  is  $n \times n$ . We likewise partition  $M$  into blocks

$$M = \begin{pmatrix} M_{\alpha\alpha} & M_{\alpha\beta} \\ M_{\beta\alpha} & M_{\beta\beta} \end{pmatrix} \text{ where } M_{\alpha\alpha} \text{ is } n \times n.$$

Thus,

$$\begin{aligned} & \kappa(sX, sX) \\ &= v(X)\Sigma(s)M\Sigma(s)v(X)^t \\ &= v_\alpha(X) \begin{pmatrix} I & \chi \\ & \chi^t \end{pmatrix} \Sigma(s)M\Sigma(s) \begin{pmatrix} I \\ \chi^t \end{pmatrix} v_\alpha(X)^t \\ &= v_\alpha(X)\Sigma_\alpha(s) \begin{pmatrix} I & \Sigma_\alpha(s)^{-1}\chi\Sigma_\beta(s) \\ & (\Sigma_\alpha(s)^{-1}\chi\Sigma_\beta(s))^t \end{pmatrix} M \begin{pmatrix} I \\ (\Sigma_\alpha(s)^{-1}\chi\Sigma_\beta(s))^t \end{pmatrix} \Sigma_\alpha(s)v_\alpha(X)^t \\ &= v_\alpha(X)\Sigma_\alpha(s) \begin{pmatrix} I & \tilde{\chi}(s) \\ & \tilde{\chi}(s)^t \end{pmatrix} M \begin{pmatrix} I \\ \tilde{\chi}(s)^t \end{pmatrix} \Sigma_\alpha(s)v_\alpha(X)^t \\ &= v_\alpha(X)\Sigma_\alpha(s)\tilde{M}(s)\Sigma_\alpha(s)v_\alpha(X)^t, \end{aligned}$$

where we have implicitly defined

$$\begin{aligned} \tilde{\chi}(s) &\equiv \Sigma_\alpha(s)^{-1}\chi\Sigma_\beta(s) \\ \tilde{M}(s) &\equiv \begin{pmatrix} I & \tilde{\chi}(s) \\ & \tilde{\chi}(s)^t \end{pmatrix} M \begin{pmatrix} I \\ \tilde{\chi}(s)^t \end{pmatrix}. \end{aligned}$$

For  $0 \leq i < n$ ,  $0 \leq j < \infty$ , the  $i, j$ th entry of  $\tilde{\chi}(s)$  is  $s^{|I_{j+n}| - |I_i|}\chi_{ij}$ , and  $|I_{j+n}| - |I_i| \geq 0$ . Thus,  $\lim_{s \rightarrow 0} \tilde{\chi}(s)$  exists and we denote it  $\tilde{\chi}(0)$ . We note that  $\tilde{\chi}_{ij}(0)$  is non-zero if and only if  $|I_i| = |I_{j+n}|$ . In particular,

$$\tilde{\chi}_{ij}(0) = \begin{cases} \chi_{ij} & d_{|I_n|-1} \leq i < n \text{ and } 0 \leq j < d_{|I_n|} - n \\ 0 & \text{otherwise} \end{cases}$$

Therefore,  $\lim_{s \rightarrow 0} \tilde{M}(s) = \begin{pmatrix} I & \tilde{\chi}(0) \\ & \tilde{\chi}(0)^t \end{pmatrix} M \begin{pmatrix} I \\ \tilde{\chi}(0)^t \end{pmatrix}$  exists and is positive definite (since  $\begin{pmatrix} I & \tilde{\chi}(0) \\ & \tilde{\chi}(0)^t \end{pmatrix}$  is full rank); we denote it by  $\tilde{M}(0)$ . Additionally, since the first  $d_{|I_n|-1}$  rows of  $\tilde{\chi}(0)$  (and therefore the first  $d_{|I_n|-1}$  columns of  $\tilde{\chi}(0)^t$ ) are identically zero, the  $d_{|I_n|-1} \times d_{|I_n|-1}$  upper-left submatrices of  $\tilde{M}(0)$  and  $M$  are equal.

Summarizing,

$$\begin{aligned} f_s(y) &= s^{2p}y^t \kappa(sX, sX)^{-1}y \\ &= (v_\alpha(X)^{-1}y)^t (s^p \Sigma_\alpha(1/s)) \tilde{M}(s)^{-1} (s^p \Sigma_\alpha(1/s)) (v_\alpha(X)^{-1}y) \\ &= \tilde{q}^t (s^p \Sigma_\alpha(1/s)) \tilde{M}(s)^{-1} (s^p \Sigma_\alpha(1/s)) \tilde{q}, \end{aligned}$$

where  $\tilde{q} \equiv v_\alpha(X)^{-1}y$ . We will take the limit by applying Corollary 10.

Consider first the situation where  $p \in \mathbb{Z}$ . The first  $d_{p-1}$  diagonal entries are of the form  $s^k$  for  $k > 0$ , the ‘‘middle’’  $d_{p-1} - d_p$  entries are exactly 1, and the last  $n - d_p$  diagonal entries are of the

form  $s^{-k}$  for  $k > 0$ . We define three subsets of  $\{0, \dots, n-1\}$  (with subvectors and submatrices defined accordingly):  $lo = \{0, \dots, d_{p-1} - 1\}$ ,  $mi = \{d_{p-1}, \dots, d_p - 1\}$ , and  $hi = \{d_p, \dots, n-1\}$ . (Note it is possible for one of  $lo$  or  $hi$  to be empty, in (respectively) the cases where  $p = 0$  or  $d_p = n$ .) By Corollary 10, with  $q_1 = \tilde{q}_{lo}$ ,  $q_2 = \tilde{q}_{mi}$ , and  $q_3 = \tilde{q}_{hi}$ ,  $Z_1(s) = (s^p \Sigma_\alpha(1/s))_{lo,lo}$ , and  $Z_2^{-1}(s) = (s^p \Sigma_\alpha(1/s))_{hi,hi}$ , and

$$\begin{pmatrix} A(s) & B(s) & C(s) \\ B(s)^t & D(s) & E(s) \\ C(s)^t & E(s)^t & F(s) \end{pmatrix} = \begin{pmatrix} \tilde{M}(s)_{lo,lo}^{-1} & \tilde{M}(s)_{lo,mi}^{-1} & \tilde{M}(s)_{lo,hi}^{-1} \\ \tilde{M}(s)_{mi,lo}^{-1} & \tilde{M}(s)_{mi,mi}^{-1} & \tilde{M}(s)_{mi,hi}^{-1} \\ \tilde{M}(s)_{hi,lo}^{-1} & \tilde{M}(s)_{hi,mi}^{-1} & \tilde{M}(s)_{hi,hi}^{-1} \end{pmatrix}.$$

By Lemma 16

$$\begin{aligned} D(0) - E(0)^t F(0)^{-1} E(0) &= \tilde{M}(0)_{mi,mi}^{-1} - (\tilde{M}(0)_{hi,mi}^{-1})^t (\tilde{M}(0)_{hi,hi}^{-1})^{-1} \tilde{M}(0)_{hi,mi}^{-1} \\ &= (\tilde{M}(0)_{mi,mi} - \tilde{M}(0)_{mi,lo} \tilde{M}(0)_{lo,lo}^{-1} \tilde{M}(0)_{lo,mi})^{-1} \\ &= (M_{mi,mi} - M_{mi,lo} M_{lo,lo}^{-1} M_{lo,mi})^{-1}, \end{aligned}$$

where the final equality is the result of the  $d_{|I_n|-1} \times d_{|I_n|-1}$  upper-left submatrices of  $\tilde{M}(0)$  and  $M$  are equal, shown above.

By Corollary 10, we have that  $\lim_{s \rightarrow 0} \tilde{q}^t (s^p \Sigma_\alpha(1/s)) \tilde{M}(s)^{-1} (s^p \Sigma_\alpha(1/s)) \tilde{q}$  is  $\infty$  if  $(hi \neq \emptyset)$  and  $\tilde{q}_{hi} \neq 0$ . If  $(hi = \emptyset)$  or  $\tilde{q}_{hi} = 0$ , the limit is:

$$\tilde{q}_{mi}^t (M_{mi,mi} - M_{mi,lo} M_{lo,lo}^{-1} M_{lo,mi})^{-1} \tilde{q}_{mi},$$

hence  $f_s(y) \rightarrow R_p^k(q)$  for  $p \in \mathbb{Z}$ .

When  $p \notin \mathbb{Z}$ , the proof proceeds along very similar lines; we merely point out that in this case, we will take  $lo = \{0, \dots, d_{\lfloor p \rfloor} - 1\}$ ,  $mi = \emptyset$ , and  $hi = \{d_{\lfloor p \rfloor}, \dots, n-1\}$ . Since  $mi$  is empty, the application of Corollary 10 yields 0 when  $\tilde{q}_{hi} = 0$ , and  $\infty$  otherwise. ■

The proof assumes  $p \in [0, |I_{n-1}|)$ . In other words, we can get polynomial behavior of degree  $\lfloor p \rfloor$  for any  $p$ , but we must have at least  $d_{\lfloor p \rfloor} = O(d^{\lfloor p \rfloor})$  generic data points in order to do so.

We have shown that if  $\lambda(s) = s^{2p}$  for a  $p$  in a suitable range, that the regularization term approaches the indicator function for polynomials of degree  $p$  in the data with (when  $p \in \mathbb{Z}$ ) a residual regularization on the degree  $p$  monomial coefficients which is a quadratic form given by some combination of the coefficients of the power series expansion of  $\kappa(x, x')$ . Obtaining these coefficients in general may be awkward. However, for kernels which satisfy Lemma 1, this can be done easily.

**Lemma 13** *If  $\kappa$  satisfies the conditions of Lemma 1, then for  $p \in \mathbb{Z}$  and  $q(x) = \sum_{|I| \leq p} \tilde{q}_I x^I$*

$$R_p^k(q) = (g_p(0))^{-2} \sum_{|I|=p} \begin{pmatrix} |I| \\ I \end{pmatrix}^{-1} q_I^2. \tag{9}$$

**Proof** Let  $L$  be defined according to the proof of Lemma 1. Lemma 17 applies with  $G$  and  $J$  being the consecutive  $d_{p-1} \times d_{p-1}$  and  $(d_p - d_{p-1}) \times (d_p - d_{p-1})$  diagonal blocks of  $L$ . Finally, we note that  $J$  is itself a diagonal matrix and hence,  $(JJ^t)^{-1}$  is diagonal with elements equal to the inverse squares of  $J$ 's, i.e. of the form  $\begin{pmatrix} |I| \\ I \end{pmatrix}^{-1} (g_{|I|}(0))^{-2}$  where  $|I| = p$ . ■

Note, the Gaussian kernel is of this sort with  $(g_p(0))^2 = \frac{1}{p!}$ .

It is also worth noting that for kernels admitting such a decomposition  $R_p^{\kappa}(q)$  is invariant under “rotations” of the form  $q \rightarrow q'$  where  $q(x) = q'(Ux)$  with  $U$  a rotation matrix. Since any  $R_p^{\kappa}(q) = 0$  for  $q$  of degree  $< p$  it is clearly translation invariant. We speculate that any quadratic function of the coefficients of a polynomial which is both translation and rotation invariant in this way must have of the form (9).

## 6. The Asymptotic Regularized Solution

By Theorem 12, the regularization term (under certain conditions) becomes a penalty on degree  $> p$  behavior of the regularized solution. Since the loss function is fixed as  $\sigma, \frac{1}{\lambda} \rightarrow \infty$ , the objective function in (1) approaches a limiting constrained optimization problem.

**Theorem 14** *Let  $v: \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$  be a valid loss function and  $\kappa(x, x')$  be a valid kernel function. Let  $\kappa_{\sigma}(x, x') = \kappa(\sigma^{-1}x, \sigma^{-1}x')$ . Let  $p \in [0, |I_{n-1}|)$  with  $\lambda(\sigma) = \tilde{\lambda}\sigma^{-2p}$  for some fixed  $\tilde{\lambda} > 0$ .*

*Let  $\hat{f}_{\sigma}, \hat{f}_{\infty} \in \mathcal{H}$  be the unique minimizers of*

$$n\lambda(\sigma)\|f\|_{\kappa_{\sigma}}^2 + \sum_{i=1}^n v(f(x_i), \hat{y}_i) \quad (10)$$

and

$$n\tilde{\lambda}R_p^{\kappa}(f) + \sum_{i=1}^n v(f(x_i), \hat{y}_i) \quad (11)$$

respectively.

Then  $\forall x_0 \in \mathbb{R}^d$  such that  $X_0 = \begin{pmatrix} x_0 \\ X \end{pmatrix}$  is generic,

$$\lim_{\sigma \rightarrow \infty} \hat{f}_{\sigma}(x_0) = \hat{f}_{\infty}(x_0).$$

**Proof** In the value-based learning formulation, problem 10 becomes

$$n\lambda(\sigma)y^t K_{\sigma}^{-1}y + \sum_{i=1}^n v(y_i, \hat{y}_i) \quad (12)$$

where  $y \in \mathbb{R}^n$ .

By Corollary 3, if we consider the expanded problem which includes the test point in the regularization but not in the loss,

$$n\lambda(\sigma)z^t \begin{pmatrix} \kappa(x_0, x_0) & k_{\sigma}^t \\ k_{\sigma} & K_{\sigma} \end{pmatrix}^{-1} z + \sum_{i=1}^n v(z_i, \hat{y}_i), \quad (13)$$

then the minimizers of problems 12 and 13 are related via  $\dot{z}_{\sigma i} = \dot{y}_{\sigma i} = \dot{f}_{\sigma}(x_i)$ ,  $1 \leq i \leq n$  and  $\dot{z}_{\sigma 0} = k_{\sigma}K_{\sigma}^{-1}\dot{y}_{\sigma} = \dot{f}_{\sigma}(x_0)$ . Because  $X_0$  is generic, we can make the change of variables  $z_i = q(x_i) = \sum_{j=0}^n \beta_j x_i^j$  in (13), yielding

$$g_{\sigma}(q) = n\lambda(\sigma)\|q\|_{\kappa_{\sigma}}^2 + \sum_{i=1}^n v(q(x_i), \hat{y}_i) \quad (14)$$

with minimizer  $\dot{q}_\sigma$  satisfying  $\dot{q}_\sigma(x_i) = \dot{z}_{\sigma i}$  (in particular  $\dot{q}_\sigma(x_0) = \dot{z}_{\sigma 0} = \dot{f}_\sigma(x_0)$ ).

Let  $g_\infty(q) = n\tilde{\lambda}R_p^K(q) + \sum_{i=1}^n v(q(x_i), \hat{y}_i)$  with minimizer  $\dot{q}_\infty$ . By Theorem 12,  $g_\sigma \rightarrow g_\infty$ , thus, by Theorem 7,  $\dot{q}_\sigma(x_0) \rightarrow \dot{q}_\infty(x_0) = \dot{f}_\infty(x_0)$ . ■

We note that in Theorem 14, we have assumed that problems 10 and 11 have unique minimizers. For any fixed  $\sigma$ ,  $\|f\|_{\kappa_\sigma}^2$  is strictly convex, so problem 10 will always have a unique minimizer. For strictly convex loss functions, such as the square loss used in regularized least squares, problem 11 will have a unique minimizer as well. If we consider a non-strictly convex loss function, such as the hinge loss used in SVMs, problem 11 may not have a unique minimizer; for example, it is easy to see that in a classification task where the data is *separable* by a degree  $p$  polynomial, any (appropriately scaled) degree  $p$  polynomial that separates the data will yield an optimal solution to problem 11 with cost 0. In these cases, Theorem 12 still determines the *value* of the limiting solution, but Theorem 14 does not completely determine the limiting minimizer. Theorem 7.33 of Rockafellar and Wets (2004) provides a generalization of Theorem 14 which applies when the minimizers are non-unique (and even when the objective functions are non-convex, as long as certain *local convexity* conditions hold). It can be shown that the minimizer of problem 10 will converge to one of the minimizers of problem 11, though not knowing which one, we cannot predict the limiting regularized solution. In practice, we expect that when the data is not separable by a low-degree polynomial (most real-world data sets are not), problem 11 will have a unique minimizer.

Additionally, we note that our work has focused on “standard” Tikhonov regularization problems, in which the function  $f$  is “completely” regularized. In practice, the SVM (for reasons that we view as largely historical, although that is beyond the scope of this paper) is usually implemented with an unregularized bias term  $b$ . We point out that our main result still applies. In this case,

$$\begin{aligned} & \inf_{b \in \mathbb{R}, f \in \mathcal{H}} \left\{ n\lambda \|f\|_{\kappa_\sigma} + \sum_{i=1}^n (1 - (f(x_i) + b)\hat{y}_i)_+ \right\} \\ &= \inf_b \left\{ \inf_f \left\{ n\lambda \|f\|_{\kappa_\sigma} + \sum_{i=1}^n (1 - (f(x_i) + b)\hat{y}_i)_+ \right\} \right\} \\ &\rightarrow \inf_b \left\{ \inf_f \left\{ n\tilde{\lambda}R_p^K(f) + \sum_{i=1}^n (1 - (f(x_i) + b)\hat{y}_i)_+ \right\} \right\}, \end{aligned}$$

with our results applying to the inner optimization problem (where  $b$  is fixed). When an unregularized bias term is used, problem 10 may not have a unique minimizer either. The conditions for non-uniqueness of 10 for the case of support vector machines are explored in Burges and Crisp (1999); the conditions are fairly pathological, and SVMs nearly always have unique solutions in practice. Finally, we note that the limiting problem is one where all polynomials of degree  $< p$  are free, and hence, the bias term is “absorbed” into what is already free in the limiting problem.

## 7. Prior Work

We are now in a position to discuss in some detail the previous work on this topic.

In Keerthi and Lin (2003), it was observed that SVMs with Gaussian kernels produce classifiers which approach those of linear SVMs as  $\sigma \rightarrow \infty$  (and  $\frac{1}{2\lambda} = C = \tilde{C}\sigma^2 \rightarrow \infty$ ). The proof is based on an expansion of the kernel function (Equation 2.8 from Keerthi and Lin (2003)):

$$\kappa_\sigma(x, x') = \exp(-\|x - x'\|^2 / \sigma^2)$$

$$= 1 - \frac{\|x\|^2}{2\sigma^2} - \frac{\|x'\|^2}{2\sigma^2} + \frac{x \cdot x'}{\sigma^2} + o(\|x - x'\|/\sigma^2)$$

where  $\kappa_\sigma$  is approximated by the four leading terms in this expansion. This approximation ( $\kappa_\sigma(x, x') \sim 1 - \sigma^{-2}(\|x\|^2 - \|x'\|^2 + 2x \cdot x')/2$ ) does not satisfy the Mercer condition, so the resulting dual objective function is not positive definite (remark 3 of Keerthi and Lin (2003)). However, by showing that the domain of the dual optimization problem is bounded (because of the dual box constraints), one avoids the unpleasant effects of the Mercer violation. The Keerthi and Lin (2003) result is a special case of our result, where we choose the Gaussian loss function and  $p = 1$ .

In Lippert and Rifkin (2006), a similar observation was made in the case of Gaussian regularized least squares. In this case, for any degree  $p$ , an asymptotic regime was identified in which the regularized solution approached the least squares degree- $p$  polynomial. The result hinges upon the simultaneous cancellation effects between the coefficients  $c(\sigma, \lambda)$  and the kernel function  $\kappa_\sigma$  in the kernel expansion of  $f(x)$ , with  $f(x)$  and  $c(\sigma, \lambda)$  given by

$$\begin{aligned} f(x) &= \sum_i c_i(\sigma, \lambda) \kappa_\sigma(x, x_i) \\ c(\sigma, \lambda) &= (\kappa_\sigma(X, X) + n\lambda I)^{-1} y \end{aligned}$$

when  $\kappa_\sigma(x, x') = \exp(-\|x - x'\|^2/\sigma^2)$ . In that work, we considered only *non-integer*  $p$ , so there was no residual regularization. The present work generalizes the result to arbitrary  $p$  and arbitrary convex loss-functions. Note that in our previous work, we did not work with the value-based formulation of learning, and we were forced to take the limit of an expression combining training and testing kernel products, exploiting the explicit nature of the regularized least squares equations. In the present work, the value-based learning formulation allows us to avoid such issues, obtaining much more general results.

## 8. Experimental Evidence

In this section, we present a simple experiment that illustrates our results. This example was first presented in our earlier work (Lippert and Rifkin, 2006).

We consider the fifth-degree polynomial function

$$f(x) = .5(1 - x) + 150x(x - .25)(x - .3)(x - .75)(x - .95),$$

over the range  $x \in [0, 1]$ . Figure 3 plots  $f$ , along with a 150 point data set drawn by choosing  $x_i$  uniformly in  $[0, 1]$ , and choosing  $y = f(x) + \epsilon_i$ , where  $\epsilon_i$  is a Gaussian random variable with mean 0 and standard deviation .05. Figure 3 also shows (in red) the best polynomial approximations to the data (not to the ideal  $f$ ) of various orders. (We omit third order because it is nearly indistinguishable from second order.)

According to Theorem 14, if we parametrize our system by a variable  $s$ , and solve a Gaussian regularized least-squares problem with  $\sigma^2 = s^2$  and  $\lambda = \tilde{\lambda}s^{-(2p+1)}$  for some integer  $p$ , then, as  $s \rightarrow \infty$ , we expect the solution to the system to tend to the  $p$ th-order data-based polynomial approximation to  $f$ . Asymptotically, the value of the constant  $\tilde{\lambda}$  does not matter, so we (arbitrarily) set it to be 1. Figure 4 demonstrates this result.

We note that these experiments frequently require setting  $\lambda$  much smaller than machine- $\epsilon$ . As a consequence, we need more precision than IEEE double-precision floating-point, and our results

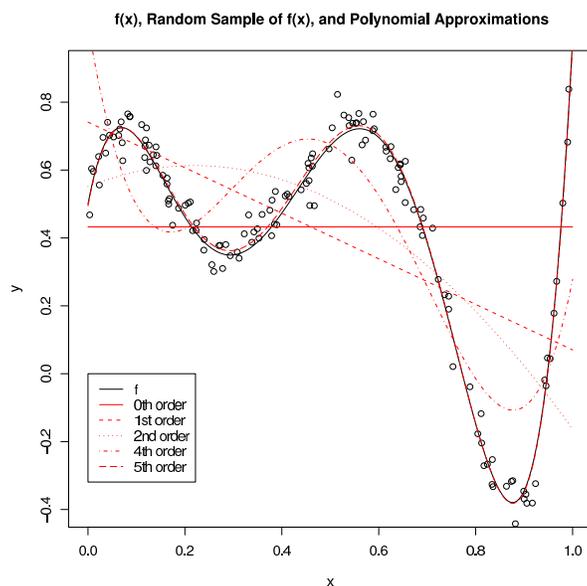


Figure 3:  $f(x) = .5(1-x) + 150x(x-.25)(x-.3)(x-.75)(x-.95)$ , a random data set drawn from  $f(x)$  with added Gaussian noise, and data-based polynomial approximations to  $f$ .

cannot be obtained via many standard tools (e.g. MATLAB(TM)). We performed our experiments using CLISP, an implementation of Common Lisp that includes arithmetic operations on arbitrary-precision floating point numbers.

## 9. Discussion

We have shown, under mild technical conditions, that the minimizer of a Tikhonov regularization problem with a Gaussian kernel with bandwidth  $\sigma$  behaves, as  $\sigma \rightarrow \infty$  and  $\tilde{\lambda} = \tilde{\lambda}\sigma^{-p}$ , like the degree- $p$  polynomial that minimizes empirical risk (with some additional regularization on the degree  $p$  coefficients when  $p$  is an integer). Our approach rested on two key ideas, epi-convergence, which allowed us to make precise statements about when the limits of minimizers converges to the minimizer of a limit, and value-based learning, which allowed us to work in terms of the predicted functional values,  $y_i$ , as opposed to the more common technique of working with the coefficients  $c_i$  in a functional expansion of the form  $f(x) = \sum_i c_i K(x, x_i)$ . This in turn allowed us to avoid discussing the limits of the  $c_i$ , which we do not know how to characterize.

We are *not* suggesting that practitioners wishing to do polynomial approximation use Gaussian kernels with extreme  $\sigma, \lambda$  values; there is no difficulty in using standard polynomial kernels directly, and using extreme  $\sigma$  and  $\lambda$  values invites numerical difficulties. However, we think this result highlights a phenomenon which may mislead automated parameter tuning methods (such as selecting  $\sigma$  or  $\lambda$  to minimize some hold-out error). In fact, our earlier work (Lippert and Rifkin, 2006), was motivated by experiments in globally optimizing the LOO error in  $(\lambda, \sigma)$ , where, for some data sets we observed large ranges of decreasing  $\lambda$  and increasing  $\sigma$  which had similar, nearly optimal performance. Wahba et al. (2001) observed the same phenomenon for the SVM, optimizing

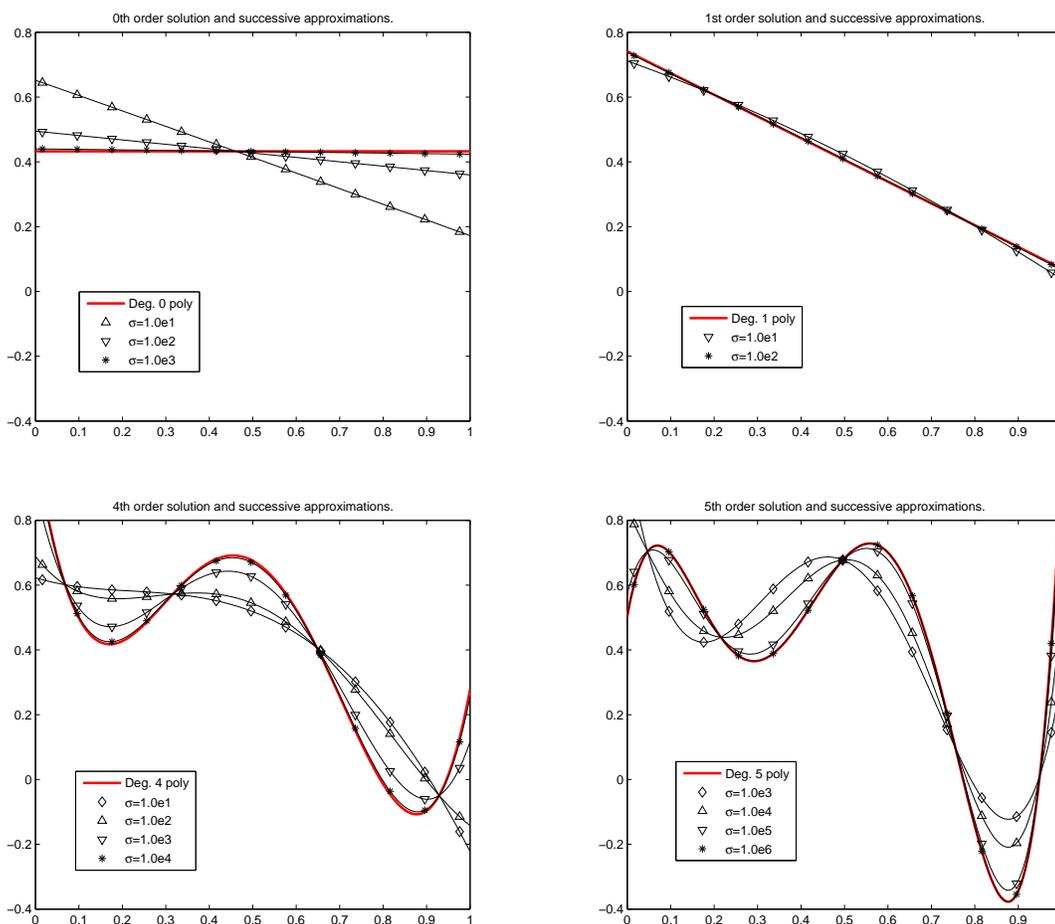


Figure 4: As  $s \rightarrow \infty$ ,  $\sigma^2 = s^2$  and  $\lambda = s^{-(2k+1)}$ , the solution to Gaussian RLS approaches the  $k$ th order polynomial solution.

performance of a Bayesian weighted misclassification score. One can get some intuition about this tradeoff between smaller  $\lambda$  and larger  $\sigma$  by considering example 4 of Zhou (2002) where a tradeoff between  $\sigma$  and  $R$  is seen for the covering numbers of balls in an RKHS induced by a Gaussian kernel ( $R$  can be thought of as roughly  $\frac{1}{\sqrt{\lambda}}$ ).

We think it is interesting that some low-rank approximations to Gaussian kernel matrix-vector products (see Yang et al. (2005)) tend to work much better for large values of  $\sigma$ . Our results raise the possibility that these low-rank approximations are merely recovering low-order polynomial behavior; this will be a topic of future study.

We believe the value-based formulation is of quite general utility, and expect to work with it in the future. Because of our choice of kernels, we were able to assume that the kernel matrix  $K$  was invertible, and we worked directly with  $K^{-1}$  in the value-based formulation. This is not a strong

requirement; it is possible to work with the pseudoinverse of  $K$  for finite-dimensional kernels (such as the dot-product kernel).

## Acknowledgments

The authors would like to acknowledge Roger Wets and Adrian Lewis for patiently answering questions regarding epi-convergence. We would also like to thank our reviewers for their comments and suggestions.

## Appendix A

In this appendix, we state and prove several matrix identities that we use in the main body of the paper.

**Lemma 15** *Let  $X, U \in \mathbb{R}^{m \times m}$ ,  $Z, W \in \mathbb{R}^{n \times n}$ , and  $Y, V \in \mathbb{R}^{n \times m}$  with  $\begin{pmatrix} X & Y^t \\ Y & Z \end{pmatrix}$  symmetric, positive definite. If*

$$\begin{pmatrix} U & V^t \\ V & W \end{pmatrix} \begin{pmatrix} X & Y^t \\ Y & Z \end{pmatrix} = \begin{pmatrix} I_m & 0 \\ 0 & I_n \end{pmatrix} \quad (15)$$

then

$$U = (X - Y^t Z^{-1} Y)^{-1} \quad (16)$$

$$W^{-1} V = -Y X^{-1} \quad (17)$$

$$V U^{-1} = -Z^{-1} Y \quad (18)$$

$$W = (Z - Y X^{-1} Y^t)^{-1} \quad (19)$$

**Proof** Since  $\begin{pmatrix} X & Y^t \\ Y & Z \end{pmatrix}$ , is symmetric, positive definite,  $\begin{pmatrix} U & V^t \\ V & W \end{pmatrix}$  is symmetric, positive definite, as are  $X, Z, U, W$ .

Multiplying out (15) in block form,

$$U X + V^t Y = I_m \quad (20)$$

$$V X + W Y = 0 \quad (21)$$

$$U Y^t + V^t Z = 0 \quad (22)$$

$$V Y^t + W Z = I_n \quad (23)$$

Since  $U, W, X, Z$  are non-singular, (21) implies (17) and (22) implies (18). Substituting  $V = -Z^{-1} Y U$  into (20) yields  $U X - U Y^t Z^{-1} Y = U(X - Y^t Z^{-1} Y) = I_m$  and thus (16). Similarly,  $V = -W Y X^{-1}$  and (23) give (19). ■

**Lemma 16** *Let*

$$M = \begin{pmatrix} A & B^t & C^t \\ B & D & E^t \\ C & E & F \end{pmatrix}$$

be symmetric positive definite. Let

$$M^{-1} = \begin{pmatrix} \bar{A} & \bar{B}^t & \bar{C}^t \\ \bar{B} & \bar{D} & \bar{E}^t \\ \bar{C} & \bar{E} & \bar{F} \end{pmatrix}.$$

Then  $\bar{D} - \bar{E}^t \bar{F}^{-1} \bar{E} = (D - BA^{-1}B^t)^{-1}$ .

**Proof** By (16) of Lemma 15, on  $M$  with  $U = \begin{pmatrix} A & B^t \\ B & D \end{pmatrix}$ ,

$$\begin{pmatrix} A & B^t \\ B & D \end{pmatrix}^{-1} = \begin{pmatrix} \bar{A} & \bar{B}^t \\ \bar{B} & \bar{D} \end{pmatrix} - \begin{pmatrix} \bar{C}^t \\ \bar{E}^t \end{pmatrix} \bar{F}^{-1} (\bar{C} \ \bar{E}).$$

By (19) of Lemma 15, on  $\begin{pmatrix} A & B^t \\ B & D \end{pmatrix}$ ,

$$\begin{pmatrix} A & B^t \\ B & D \end{pmatrix}^{-1} = \begin{pmatrix} \cdots & \cdots \\ \cdots & (D - BA^{-1}B^t)^{-1} \end{pmatrix}.$$

Combining the lower-right blocks of the above two expansions yields the result. ■

**Lemma 17** *If*

$$M = \begin{pmatrix} A & B^t & C^t \\ B & D & E^t \\ C & E & F \end{pmatrix} = \begin{pmatrix} G & 0 & 0 \\ H & J & 0 \\ K & N & P \end{pmatrix} \begin{pmatrix} G & 0 & 0 \\ H & J & 0 \\ K & N & P \end{pmatrix}^t.$$

*is symmetric positive definite, then  $JJ^t = D - BA^{-1}B^t$ .*

**Proof** Clearly  $\begin{pmatrix} A & B^t \\ B & D \end{pmatrix} = \begin{pmatrix} G & 0 \\ H & J \end{pmatrix} \begin{pmatrix} G & 0 \\ H & J \end{pmatrix}^t$  and thus

$$A = GG^t, \quad B = HG^t, \quad D = JJ^t + HH^t$$

and hence  $JJ^t = D - HH^t = D - BG^{-t}G^{-1}B^t = D - B(GG^t)^{-1}B^t = D - BA^{-1}B^t$ . ■

## References

- N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68:337–404, 1950.
- C. J. C. Burges and D. J. Crisp. Uniqueness of the svm solution. In *Neural Information Processing Systems*, 1999.
- Theodoros Evgeniou, Massimiliano Pontil, and Tomaso Poggio. Regularization networks and support vector machines. *Adv. In Comp. Math.*, 13(1):1–50, 2000.
- Federico Girosi, Michael Jones, and Tomaso Poggio. Regularization theory and neural networks architectures. *Neural Computation*, 7(2):219–269, 1995.

- S. Sathya Keerthi and Chih-Jen Lin. Asymptotic behaviors of support vector machines with Gaussian kernel. *Neural Computation*, 15(7):1667–1689, 2003.
- Ross A. Lippert and Ryan M. Rifkin. Asymptotics of gaussian regularized least squares. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Adv. in Neural Info. Proc. Sys. 18*. MIT Press, Cambridge, MA, 2006.
- R. Tyrrell Rockafellar and Roger J. B. Wets. *Variational Analysis*. Springer, Berlin, 2004.
- Bernhard Schölkopf, Ralf Herbrich, and Alex J. Smola. A generalized representer theorem. In *14th Annual Conference on Computational Learning Theory*, pages 416–426, 2001.
- Grace Wahba. *Spline Models for Observational Data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Soc. for Industrial & Appl. Math., 1990.
- Grace Wahba, Yi Lin, Yoonkyung Lee, and Hao Zhang. On the relation between the GACV and Joachims’  $\xi\alpha$  method for tuning support vector machines, with extensions to the non-standard case. Technical Report 1039, U. Wisconsin department of Statistics, 2001. URL [citeseer.ist.psu.edu/wahba01relation.html](http://citeseer.ist.psu.edu/wahba01relation.html).
- Changjiang Yang, Ramani Duraiswami, and Larry Davis. Efficient kernel machines using the improved fast gauss transform. In Lawrence K. Saul, Yair Weiss, and Léon Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 1561–1568, Cambridge, MA, 2005. MIT Press.
- D.-X. Zhou. The covering number in learning theory. *J. of Complexity*, 18:739–767, 2002.