Journal of Machine Learning Research 8 (2007) 441-479 Submitted 3/06; Revised 10/06; Published 3/07

Value Regularization and Fenchel Duality

Ryan M. Rifkin RRIFKIN@HONDA-RI.COM
Honda Research Institute USA, Inc.
One Cambridge Center, Suite 401
Cambridge, MA 02142, USA

Ross A. Lippert LIPPERT@MATH.MIT.EDU
Department of Mathematics
Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge, MA 02139-4307, USA

Editor: Peter Bartlett

Abstract

Regularization is an approach to function learning that balances fit and smoothness. In practice, we search for a function f with a finite representation $f = \sum_i c_i \phi_i(\cdot)$. In most treatments, the c_i are the primary objects of study. We consider *value regularization*, constructing optimization problems in which the predicted values at the training points are the primary variables, and therefore the central objects of study. Although this is a simple change, it has profound consequences. From convex conjugacy and the theory of Fenchel duality, we derive separate optimality conditions for the regularization and loss portions of the learning problem; this technique yields clean and short derivations of standard algorithms. This framework is ideally suited to studying many other phenomena at the intersection of learning theory and optimization. We obtain a value-based variant of the representer theorem, which underscores the transductive nature of regularization in reproducing kernel Hilbert spaces. We unify and extend previous results on learning kernel functions, with very simple proofs. We analyze the use of unregularized bias terms in optimization problems, and low-rank approximations to kernel matrices, obtaining new results in these areas. In summary, the combination of value regularization and Fenchel duality are valuable tools for studying the optimization problems in machine learning.

Keywords: kernel machines, duality, optimization, convex analysis, kernel learning

1. Introduction

Given a set of training data $\{(X_1,Y_1),\ldots,(X_n,Y_n)\}$, the inductive supervised learning task is to learn a function f that, given a new X value, will predict the associated Y value. A common framework for solving this problem is Tikhonov regularization (Tikhonov and Arsenin, 1977):

$$\inf_{f\in\mathcal{F}}\left\{\sum_{i=1}^{n} v(f(X_i),Y_i)+\frac{\lambda}{2}\Omega(f)\right\}. \quad (1)$$

In this general form, $\mathcal{F}$ is a space of functions from which f must be selected, and $v(f(X_i),Y_i)$ is the *loss*, indicating the penalty we pay when we see X_i, predict $f(X_i)$, and the true value is Y_i. For a large class of functions $\mathcal{F}$, simply minimizing $\sum_{i=1}^{n} v(f(X_i),Y_i)$ directly is ill-posed and leads to overfitting the training data. We restore well-posedness by introducing a *regularization* term $\Omega(f)$

that penalizes elements of f that are too "complex". The regularization parameter, λ, controls the tradeoff between finding a function of low complexity and fitting the training data.

A large amount of work (Wahba, 1990; Evgeniou et al., 2000) takes $\mathcal{F}$ to be a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ (Aronszajn, 1950) induced by a kernel function k. In this situation, we rewrite (1) as

$$\inf_{f\in\mathcal{H}} \left\{ \sum_{i=1}^{n} v(f(X_i), Y_i) + \frac{\lambda}{2}||f||_k^2 \right\}, \tag{2}$$

indicating that the regularization term is now the squared norm of the function in the RKHS. Many common algorithms, including support vector machines for classification (Cortes and Vapnik, 1995) and regression (Vapnik, 1998), regularized least squares (Wahba, 1990; Poggio and Girosi, 1990; Rifkin, 2002), and kernel logistic regression (Jaakkola and Haussler, 1999) are instances of Tikhonov regularization: different loss functions yield different algorithms.[1]

A consequence of the so-called *representer theorem* (Wahba, 1990; Schölkopf et al., 2001), is that the minimizer of (2) will have the form

$$f(x) = \sum_{i=1}^{n} c_i k(x, X_i). \tag{3}$$

In other words, in an RKHS, a function f which minimizes (2) is a sum of kernel functions $k(\cdot, X_i)$ over the data points. We solve a Tikhonov regularization (or, equivalently, train a predictor) by finding the coefficients c_i. Assuming that the loss function is convex, the minimization of (2) (or (1)) is tractable.

Defining K as the $n \times n$ *kernel matrix* with $K_{ij} = k(X_i, X_j)$, the regularization penalty $||f||_k^2$ becomes $c^t K c$, the output at training point i is given by $f(X_i) = \sum_{j=1}^{n} c_j k(X_i, X_j) = (Kc)_i$, and we can rewrite (2) as

$$\inf_{c\in\mathbb{R}^n} \left\{ \sum_{i=1}^{n} v((Kc)_i, Y_i) + \frac{\lambda}{2} c^t K c \right\}. \tag{4}$$

Regularizing in an RKHS is special in that it has a representer theorem: the optimal solution in an infinite-dimensional space of functions is found by solving a finite dimensional minimization problem. Although most other function space regularizers do not have similar properties, we may consider other regularizers as modifications of (4): replacing $c^t K c$ with $c^t c$, we obtain *ridge regression* (Tikhonov and Arsenin, 1977), and replacing it with $\sum_{i=1}^{n} |c_i|$, we obtain *L1-regularization* (Zhu et al., 2003). In such cases, we are assuming that we are looking for a function of the form (3) a priori. There is nothing wrong with this, but it should not be confused with the case of RKHS regularization, where (3) arises naturally from a search for an optimizing function.

In general, an equation like (4) is used as the starting point for thinking algorithmically about finding c. For example, in a standard development of support vector machines (Cristianini and Shaw-Taylor, 2000; Rifkin, 2002), one starts with (4) instantiated with the SVM *hinge loss* $v(f(X_i), Y_i) = (1 - y_i f(X_i))_+$, introducing slack variables $\xi_i = v(f(X_i), Y_i)$ and constraints to handle

[1] Technically speaking, to derive an algorithm such as the classic SVM one also needs an *unregularized* bias term b: this issue is discussed in detail later in the paper.

the non-differentiability of the hinge loss at 0 as well as an unregularized bias term b, yielding a quadratic program:[2]

$$\begin{array}{lll} \min\limits_{c\in\mathbb{R}^n,\xi\in\mathbb{R}^n,b\in\mathbb{R}} & \frac{1}{n}\sum_{i=1}^n \xi_i + \lambda c^T K c & \\ \text{subject to :} & Y_i\left(\sum_{j=1}^n c_j k(X_i,X_j)+b\right) \geq 1-\xi_i & i=1,\ldots,n \\ & \xi_i \geq 0 & i=1,\ldots,n. \end{array}$$

This program is called the *primal problem*. In order to expose sparsity in the solution, the Lagrangian dual is taken, yielding the *dual problem*:

$$\begin{array}{lll} \max\limits_{\alpha\in\mathbb{R}^n} & \sum_{i=1}^n \alpha_i - \frac{1}{(2\lambda)^2}\alpha^t \mathrm{diag}(Y) K \mathrm{diag}(Y)\alpha & \\ \text{subject to :} & \sum_{i=1}^n Y_i\alpha_i = 0 & \\ & 0 \leq \alpha_i \leq \frac{1}{n} & i=1,\ldots,n. \end{array}$$

It is then observed that the dual problem is easier to solve (because of its simpler constraint structure), and that solutions to the primal can be easily obtained from solutions to the dual.

We propose to take a different approach to Tikhonov regularization, that we believe to be more fundamental. The approach rests on two ideas.

First, we consider the *predicted values* $y_i \equiv f(X_i) = (Kc)_i$ to be the central objects of study, and write our optimization problems in terms of y. In the case of RKHS regularization with a positive-definite kernel, the matrix K is typically non-singular (Micchelli, 1986), [3] and we rewrite (4) as a *value regularization*:

$$\inf_{y\in\mathbb{R}^n}\left\{\frac{\lambda}{2}y^t K^{-1} y + \sum_{i=1}^n v(y_i,Y_i)\right\}.$$

Although this is a simple transformation, the consequences are far-reaching. Looking at the rewritten problem, we notice that the kernel matrix K appears only in the regularization—*it does not appear in the loss term.* This is intuitive, as the loss function (of course) cares only about the predicted outputs, not what combination of kernel coefficients generated those predicted outputs. Additionally, we see that the loss function decomposes into n separate single-point loss functions. In contrast, if the c_i are the primary variables, the loss at each data point is a function of the entire c vector.

The benefits of value regularization are greatly amplified by the second central idea of this work: instead of Lagrangian duality, we use *Fenchel duality* (Borwein and Lewis, 2000), a form of duality that is well-matched to the problems of learning theory. Although we discuss Fenchel duality in greater detail below, we present a brief overview here. Consider an optimization problem of the form:

$$\inf_{y\in\mathbb{R}^n}\{f(y)+g(y)\}. \tag{5}$$

[2] Traditional derivations parametrize the loss function instead of the regularization, with a constant $C = \frac{1}{2\lambda}$, but that is a minor point.

[3] Throughout this paper, when we work with an RKHS regularizer we will generally assume that the kernel function is positive definite and the points are in general position, implying the existence of K^{-1}. This assumption can be easily relaxed using pseudoinverses, although the mathematics becomes somewhat more cumbersome.

Fenchel duality defines a so-called Fenchel dual:

$$\sup_{z \in \mathbb{R}^n} \{-f^*(z) - g^*(-z)\}, \tag{6}$$

where f^*, g^* are *Fenchel-Legendre conjugates* (definition 4) computed via auxiliary (and by design, simpler) optimization problems on f and g *separately*. Convexity of f and g (and some topological qualifications) ensure that the optimal objective values of (5) and (6) are equivalent, and that any optimal y and z satisfy:

$$\begin{aligned} f(y) - y^t z + f^*(z) &= 0 \\ g(y) + y^t z + g^*(-z) &= 0. \end{aligned}$$

Fenchel duality encompasses other notions of duality such as Lagrangian duality and seems to be a natural concept to apply to regularization problems where f and g each arise from different considerations—for supervised learning problems, one will come from regularization and the other from empirical loss.

Elaborating on this idea, Fenchel duality gives us a *separation of concerns* which is not present in the Lagrangian approach. The point of formulating an optimization problem such as a quadratic program is ultimately to derive optimality conditions and algorithms for finding optimal solutions. For convex optimization problems, all local optima are globally optimal, and we can formulate a complete set of optimality conditions which the primal and dual solutions will simultaneously satisfy. These are generally known as the Karush-Kuhn-Tucker (KKT) conditions (Bazaraa et al., 1993).[4] Fenchel duality makes it clear that the two functions f and g, or, in our case, the loss term and the regularization, contribute *individual* local optimality conditions, and the total optimality conditions for the overall problem are simply the union of the individual optimality conditions. Put differently, using Fenchel duality, we can derive a table for n_r regularizations and n_l losses, and immediately combine these to derive optimality conditions for $n_r n_l$ learning problems. This idea is obscured by the Lagrangian duality approach to deriving optimality conditions for learning problems. As an example, for the common case of the RKHS regularizer $\frac{1}{2}y^t K^{-1} y$, we will find that the primal-dual relationship is given by $y = \lambda^{-1} K z$. This condition is *independent of the loss*—it shows up simultaneously in SVM, regularized least squares, and logistic regression.

Value regularization and Fenchel duality reinforce each other's strengths. Because the kernel affects only the regularization term and not the loss term, applying Fenchel duality to value regularization yields extremely clean formulations. The major contribution of this paper is the combination of value regularization and Fenchel duality in a framework that yields new insights into many of the optimization problems that arise in learning theory.

We present, in Section 3, a primer on convex analysis and Fenchel duality, with emphasis on the key ideas that are needed for learning theory. We believe that this section provides a sufficiently self-contained summary of convex analysis to allow a machine learning researcher to apply our framework to new problems.

In Section 4, we specialize the convex analysis results to Tikhonov value regularizations consisting of the sum of a regularization and a loss term. Section 5 specializes the result further to the RKHS case and obtains very simple derivations of well-known kernel machines.

Section 6 is concerned with value regularization in the context of L1 regularization, a regularization with the explicit goal of obtaining sparsity in the finite representation. In Section 6.1, we

[4] The complete KKT conditions for SVM can be found (among other places) in Rifkin (2002).

briefly discuss 1-norm support vector machines, and emphasize a key distinction between RKHS and other regularizations: in RKHS regularization, the c_i, which are the expansion coefficients in the finite representation of the learned function, and the z_i, which are the dual variables associated with the y_i in the value regularization problem, are identifiable ($z = \lambda^{-1}c$). This is a consequence of the RKHS regularization, and does not hold for more general regularizations. In Section 6.2, we present a simpler derivation of the relationship between support vector regression and sparse approximation, first discovered in Girosi (1998); essentially, the problems are duals, with the width of the ε-tube in the support vector regression problem becoming the sparsity regularizer in the sparse approximation problem.

In Section 7, we develop a new view of the representer theorem, in the context of value regularization. The common wisdom about the representer theorem is that it guarantees that the solution to an optimization problem in an (infinite-dimensional) function space has a finite representation as an expansion of kernel functions around the training points. While this is certainly true, we gain additional insights by considering an augmented problem in which the test points also appear in the regularization, but not in the loss. In the augmented optimization problem the predicted outputs at the training points do not change, and the expansion coefficients at the test points vanish—a representer theorem. This underscores the fact that for supervised learning in an RKHS, induction and transduction are identical. In the context of transductive or semi-supervised algorithms, the picture is more complex. There have been several recent articles that turn transductive algorithms into semi-supervised algorithms via an appeal to the representer theorem. While this is formally valid, the resulting transductive algorithm is *not* equivalent to the semi-supervised algorithm, and the transductive algorithm is perhaps the more "natural" choice. Section 7.1 is devoted to a discussion of this topic.

Recently, there has been interest in "learning the kernel". In Section 8, we may view this as a value regularization where the kernel matrix itself is an auxiliary parameter to be optimized. The value-based formulation is ideal here, because the kernel appears only in the regularization term and not in the loss term. We first derive a general result that gives optimality conditions for a general convex penalty $F(K)$ on the kernel matrix. Work to date has considered only the case where F is 0 for some set of semi-definite matrices, and infinite otherwise. Lanckriet et al. (2004) considers a case where the kernel function is a linear combination of a finite set of kernels; we will see that a minor modification to their formulation yields a representer theorem and an agreement of inductive and transductive algorithms. Argyriou et al. (2005) work with a convex set generated by an infinite, continuously parametrized set of kernel functions. We give proofs of their main results which are shorter and simpler, and also more general, allowing arbitrary convex loss functions (Argyriou et al. (2005) requires differentiability).

In Section 9, we show how infimal convolutions, a form of optimization relaxation (see Section 3.4) are useful in learning theory. We first explore the idea of "biased" regularizations, the best-known example being the unregularized bias term b in the standard formulation of support vector machines. We show that unregularized bias terms arise from infimal convolutions and that the optimality conditions implied by bias terms are independent of both the particular loss function and the particular regularization. For example, including an unregularized constant term b in any Tikhonov optimization problem yields a constraint on the dual variables $\sum z_i = 0$; this constraint does not require a particular loss function, or even that we work in an RKHS. More generally, we can include unregularized polynomial functions or even include an unregularized element of an arbitrary convex set. In Section 9.2, we see that the computation of leave-one-out values can also be viewed as a biased regularization via infimal convolution.

In Section 10, we explore low rank kernel matrix approximations such as the Nyström approximation, which consists of picking a partition $N \cup M$ of the data, and approximating K with $\tilde{K} = \begin{pmatrix} K_{NN} & K_{NM} \\ K_{MN} & K_{NM}K_{MM}^{-1}K_{MN} \end{pmatrix}$. In Willams and Seeger (2000), the authors suggest using $\tilde{K}$ in place of K. Several authors (Rifkin, 2002; Rasmusen and Williams, 2006) have found empirically that this works poorly, but that a closely related approach (the *subset of regressors* method) of constraining c_i to be zero for points not in the subset M works quite well. These two approaches are extremely closely related mathematically. With value regularization and Fenchel duality, we explore in detail the relation between these methods. The subset of regressors method is, in some sense, the "natural" algorithm arising from the low-rank matrix approximation, while the Nyström method makes an unwarranted identification of the dual variables, z, and the coefficients of expansion in the finite representation, c.

The framework presented here requires a moderate mathematical investment by the reader. For this reason, we have organized the paper so that the material on convex analysis (Section 3) can be digested in several chunks and give a roadmap (Figure 1) to illustrate which sections are accessible using only a subset of Section 3. Additionally, in Section 3.5, we provide a *cheat sheet* of key ideas from convex analysis, stated without their full qualifying conditions. Alternatively, readers may use the roadmap to skip unnecessary review.

Fenchel duality will find many uses in machine learning theory. Very recently (while the present paper was under review), Dudík and Schapire (2006) used Fenchel duality to explore maximum entropy distribution estimation under constraints, and Altun and Smola (2006) explored relations between divergence minimization and statistical inference. In summary, Fenchel duality is a powerful way to look at the regularization problems that arise in learning theory. While it requires some mathematical sophistication, the resulting formulations are elegant and yield new insights. We hope that many people will find this approach useful.

2. Notation

A training set is a set of labelled points $\{(X_1,Y_1),\ldots,(X_n,Y_n)\}$. We will sometimes use N to refer to the set $\{x_1,\ldots,x_n\}$, and we may have an additional set of (unlabelled) points of size m called M (e.g., $N = \{X_1,\ldots,X_n\}$ and $M = \{X_{n+1},\ldots,X_{n+m}\}$).

Throughout this paper, Y_i (capitalized) refers to given labels for training points. y_i are variables that we optimize over. In general, we can imagine that we are learning a function f, and $y_i = f(X_i)$, but we generally think of optimizing the y_i directly rather than using f as an intermediary.

Beginning in Section 3, we will frequently take Fenchel-Legendre conjugates and use z to denote variables conjugate to y (i.e., we are using z to refer to "dual variables"). We will also sometimes obtain functions of the form $f(x) = \sum_i c_i k(X_i,x)$ and exclusively use c_i to refer to the expansion coefficients in the finite representation of f.

We define e_i to be the n-vector whose ith entry is 1 and whose other entries are zero: the ith basis vector in the standard basis for $\mathbb{R}^n$ (n will always be clear from context). We define 1_n to be a vector of length n whose entries are all 1.

We use H to refer to affine (or hyperplane) functions, $H_{v,c}(y) = v^t y - c$. For any symmetric positive semidefinite matrix, $Q_A(y) = \frac{1}{2}y^t A y$. We write $(y)_+ = \max\{y,0\}$.

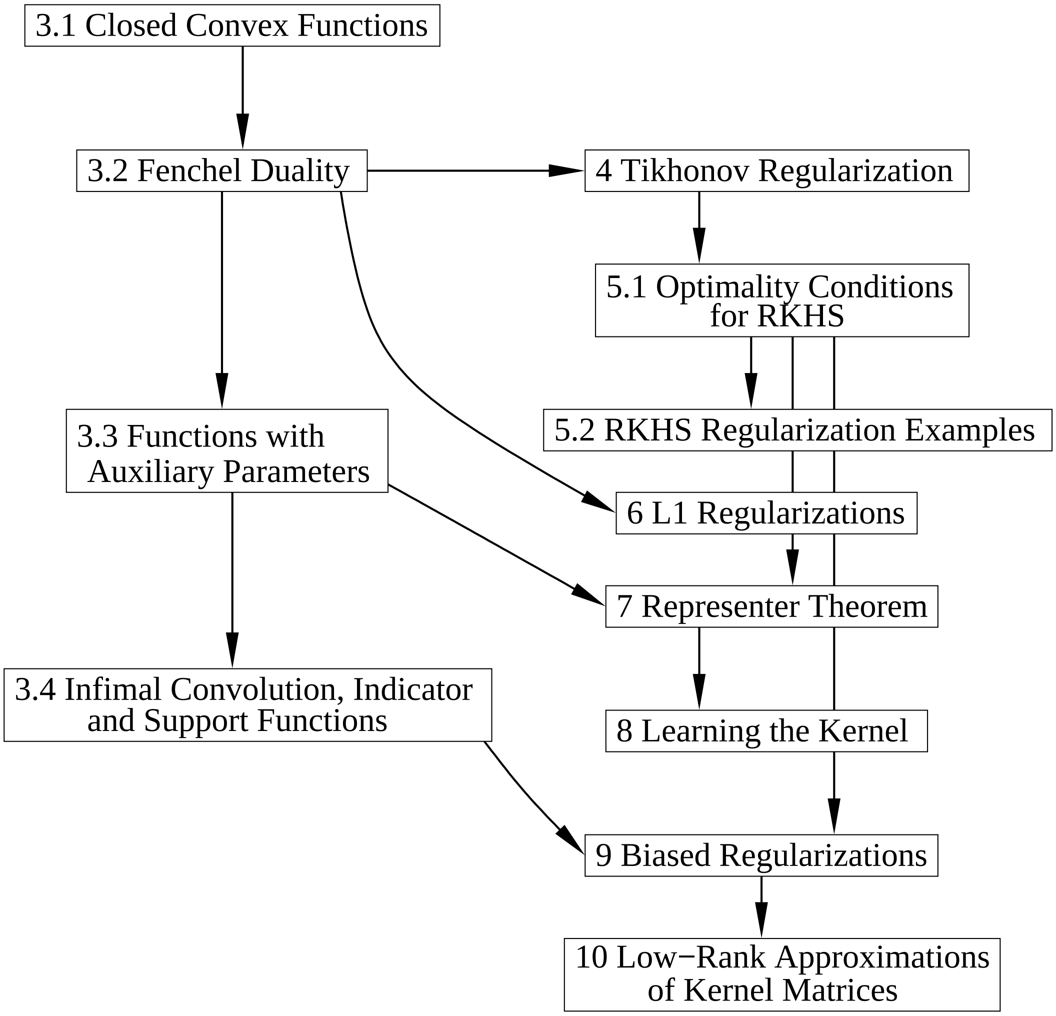
3.1 Closed Convex Functions
3.2 Fenchel Duality
4 Tikhonov Regularization
5.1 Optimality Conditions for RKHS
5.2 RKHS Regularization Examples
3.3 Functions with Auxiliary Parameters
6 L1 Regularizations
7 Representer Theorem
3.4 Infimal Convolution, Indicator and Support Functions
8 Learning the Kernel
9 Biased Regularizations
10 Low−Rank Approximations of Kernel Matrices

Figure 1: A roadmap of this paper. The sections on convex analysis are in the left-hand column, while the "applications" are in the right-hand column.

If $S, S' \subset \mathbb{R}^n$ and $A \in \mathbb{R}^{m\times n}$, then $S+S' = \{y+y' : y \in S, y' \in S'\}$, $S-S' = \{y-y' : y \in S, y' \in S'\}$, $SS' = \{yy' : y \in S, y' \in S'\}$, and $AS = \{Ay : y \in S\}$. In particular, $A\mathbb{R}^n$ is the column space of A. We denote the topological interior of $S \subset \mathbb{R}^n$ by $\text{int}(S)$. A cone is a set S with the property that $\mathbb{R}_{\geq 0}S \subset S$.

We write $\mathbb{B}_p \subset \mathbb{R}^n$ where $\mathbb{B}_p = \{y \in \mathbb{R}^n : ||y||_p \leq 1\}$ where $||\cdot||_p$ is the p-norm. We write $A^\dagger$ for the pseudoinverse of a matrix A.

3. Convex Analysis

In this section, we develop the necessary topics in convex analysis, including the needed elements of Fenchel duality theory. All results in this section can be found in Borwein and Lewis (2000) and

Rockafellar and Wets (2004), although in some cases we have substituted less general versions of the results that are sufficient for our purposes, and in some cases we have elaborated (with proofs) ideas that are introduced as exercises in these books.

3.1 Closed Convex Functions

Definition 1 *Given a function $f : \mathbb{R}^n \to [-\infty, \infty]$, the* epigraph *of f, epi f, is defined by*

$$epi\ f = \{(y, e) : e \geq f(y)\} \subset \mathbb{R}^n \times \mathbb{R}.$$

We say f is closed *or* convex *if epi f is closed, or convex.*
We define dom $f = \{y \in \mathbb{R}^n : f(y) < \infty\}$.
We say f is proper *when dom $f \neq \emptyset$ and $f > -\infty$ (i.e., $\forall y, f(y) > -\infty$).*

(Some texts consider $f : \mathbb{R}^n \to (-\infty, \infty]$, whereupon $f > -\infty$ is automatic.)

We will mostly be considering f which do not take the value $-\infty$ (such functions are somewhat pathological), in which case f being proper is equivalent to dom $f \neq \emptyset$. Allowing f to take the value of ∞ merely allows some portions of epi f to have no projection onto $\mathbb{R}^n$ (dom f is that projection). One way of viewing constrained minimization problems is as unconstrained problems with an objective function that can take the value ∞. Indicator functions (introduced in Section 3.4) are a device for this purpose. We will not do arithmetic involving ∞ except where the result is unambiguous (e.g., $\infty + 1 = \infty$).

The functions of primary interest to us are closed, convex, proper functions. We will call such a function a *ccp* function.

Definition 2 *Given $f : \mathbb{R}^n \to (-\infty, \infty]$, we define the set* $\operatorname{argmin}_{y \in \mathbb{R}^n} f(y)$ *as follows,*

$$\operatorname*{argmin}_{y \in \mathbb{R}^n} f(y) = \begin{cases} \mathbb{R}^n & \inf_{y \in \mathbb{R}^n} f(y) = \infty \\ \{y : f(y) = f_0\} & \inf_{y \in \mathbb{R}^n} f(y) = f_0 \in \mathbb{R} \\ \emptyset & \inf_{y \in \mathbb{R}^n} f(y) = -\infty \end{cases}$$

with symmetrical definitions for argmax *when needed.*

If f is proper, then the first case cannot occur. The occurrence of the middle case (given f proper) is equivalent to f being bounded from below. However, even in the second case, $\operatorname{argmin}_y f(y)$ may still be empty.

The notion of a supporting hyperplane gives us a non-smooth generalization of the gradient, called a *subgradient*.

Definition 3 (subgradients and subdifferentials) *If $f : \mathbb{R}^n \to (-\infty, \infty]$ is convex and $y \in$ dom f, then $\phi \in \mathbb{R}^n$ is a* subgradient *of f at y if it satisfies $\phi^t z \leq f(y + z) - f(y)$ for all $z \in \mathbb{R}^n$.*

The set of all such ϕ is the subdifferential *and denoted $\partial f(y)$. By convention, $\partial f(y) = \emptyset$ if $y \notin$ dom f.*

∂f is a function ($y \to \partial f(y)$) whose values are convex sets. If f is differentiable at y, then $\partial f(y)$ contains a single point, the gradient, but not so if non-differentiable at y. It is possible for the subdifferential of a convex function to be $\emptyset$ when $y \in$ dom f (for example, $f(y) = -\sqrt{y}$ has $\partial f(0) = \emptyset$ even though $f(0) = 0$). However, $\partial f(y) \neq \emptyset$ for $y \in \operatorname{int}(\text{dom } f)$ (by Theorem 3.1.8 of Borwein and Lewis, 2000).

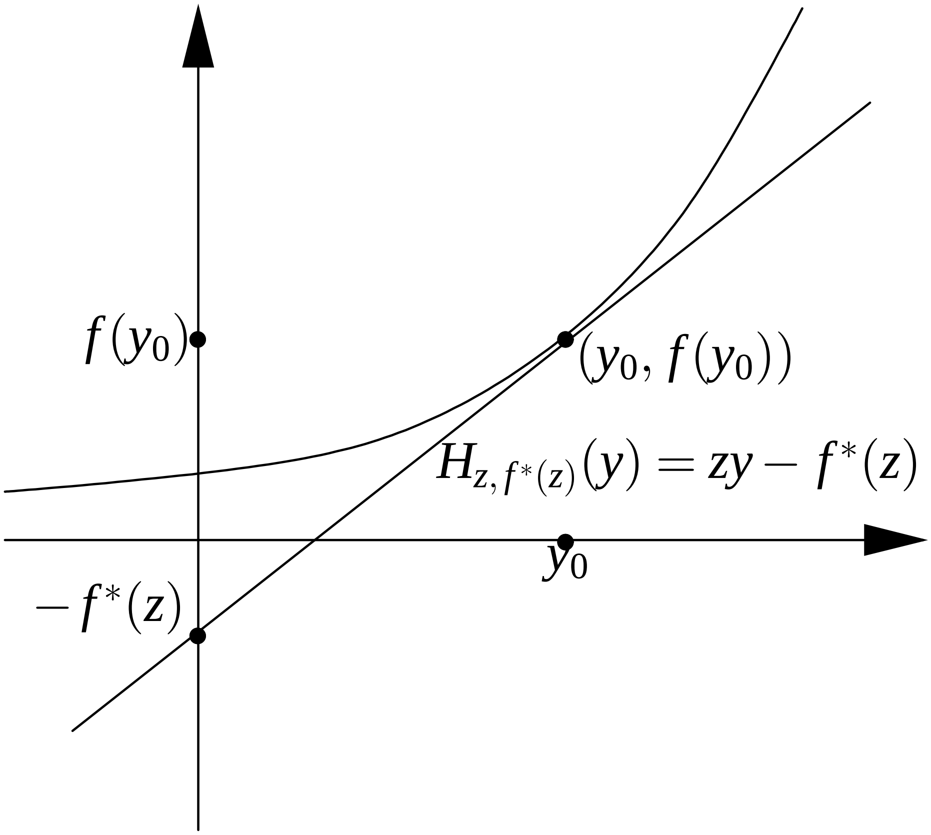

Figure 2: A graphical illustration of the Fenchel-Legendre conjugate.

3.2 Fenchel Duality

Central to Fenchel duality is the Fenchel-Legendre conjugate,

Definition 4 (Fenchel-Legendre conjugate) *Given a function* $f : \mathbb{R}^n \to [-\infty, \infty]$, *the* Fenchel-Legendre conjugate *is*

$$f^*(z) = \sup_y \{y^t z - f(y)\}. \tag{7}$$

Table 1 lists a few general, easily derived conjugation identities.

For a convex function f of one variable, one can get a sense of what the conjugate and subgradient look like by examining a graph of $f(y)$. For a given y, one finds a point $(y_0, f(y_0))$ (on the boundary of epi f) such that epi f is supported at y_0 by a line of the form $H_{z,c}(y) = zy - c$ (i.e., z is the slope of a tangent line of epi f at $(y_0, f(y_0))$). If such a supporting line exists, then $f^*(z) = c$, otherwise $f^*(z) = \infty$. Because f is convex, at most one such supporting line can exist; if f is linear in a neighborhood of y_0 then $H_{z,c}$ supports f at multiple points. Figure 2 illustrates the idea.

$f = g^*$	$g = f^*$	qualifiers
$f(y)$	$g(z)$	
$h(y)+c$	$h^*(z)-c$	
$h(y)-a^ty$	$h^*(z+a)$	
$ah(y)$	$ah^*(z/a)$	$a>0$
$h(A^{-1}y+b)$	$h^*(A^tz)-b\cdot z$	A non-singular
$\frac{1}{2}y^tAy$	$\frac{1}{2}z^tA^{-1}z$	A symmetric positive definite
v^ty+b	$\delta_{\{v\}}(z)-b$	

Table 1: Some conjugates and properties of conjugation.

More generally, we can work in terms of affine functions and epigraphs. Equation 7 is equivalent to

$$\text{epi } f^* = \bigcap_{y\in\text{dom } f} \text{epi } H_{y,f(y)}.$$

Being an arbitrary intersection of closed and convex sets, epi f^* is closed and convex. Thus, f^* is closed and convex even if f is neither. Additionally, for $z \in \text{dom } f^*$, by (7), we have $H_{z,f^*(z)}(y) = z^ty - f^*(z) \leq f(y)$ for all y, and hence,

$$\text{epi } f \subset \bigcap_{z\in\text{dom } f^*} \text{epi } H_{z,f^*(z)},$$

holding with equality when f is closed and convex (Theorem 4.2.1 of Borwein and Lewis, 2000).

Theorem 5 (biconjugation) *$f : \mathbb{R}^n \to (-\infty,\infty]$ is closed and convex iff $f^{**} = f$.*

In particular, conjugation is a bijection between ccp functions.

The supremum in (7) is attained if and only if $H_{z,f^*(z)}(y) = f(y)$ for some y. Thus, $z \in \partial f(y)$ is equivalent to $f(y) = y^tz - f^*(z)$ (Theorem 3.3.4 of Borwein and Lewis, 2000).

Theorem 6 (Fenchel-Young) *Let $f : \mathbb{R}^n \to (-\infty,\infty]$ be convex. $\forall y,z \in \mathbb{R}^n$,*

$$f(y) + f^*(z) \geq y^tz$$

with equality holding iff $z \in \partial f(y)$.

Combining the previous results, if f is closed and convex then $z \in \partial f(y) \Leftrightarrow y \in \partial f^*(z)$, and if $z \in \text{int}(\text{dom } f^*)$ the supremum in (7) is attained.

Fenchel duality can be motivated from Theorem 6 by considering two functions simultaneously. Given convex $f,g : \mathbb{R}^n \to (-\infty,\infty]$

$$f(y) + f^*(z) - y^tz \geq 0 \tag{8}$$

$$g(y) + g^*(-z) - y^t(-z) \geq 0. \tag{9}$$

Summing the above inequalities and minimizing,

$$f(y) + g(y) + f^*(z) + g^*(-z) \geq 0 \tag{10}$$

$$\inf_y\{f(y)+g(y)\} + \inf_z\{f^*(z)+g^*(-z)\} \geq 0. \tag{11}$$

If $\exists y,z \in \mathbb{R}^n$ such that (10) is an equality, then clearly (11) holds with equality as well and the individual infima are attained. Moreover, (8) and (9) become equalities and thus $z \in \partial f(y)$ and $-z \in \partial g(y)$ (and $y \in \partial f^*(z)$ and $y \in \partial g^*(-z)$, if f,g are ccp). We now quote the fundamental theorem of Fenchel duality, which supplies sufficient conditions for (11) to hold and for either of the infima to attain. Our statement is less general than Theorem 3.3.5 of Borwein and Lewis (2000), but will serve for our purposes.

Theorem 7 (Fenchel duality) *Let $f,g : \mathbb{R}^n \rightarrow (-\infty,\infty]$ be convex with $f+g$ bounded below. If $0 \in int(dom\ f - dom\ g)$, then (11) is an equality and the infimum of $\inf_z\{f^*(z)+g^*(-z)\}$ is attained.*

The topological sufficiency condition, $0 \in \text{int}(\text{dom } f - \text{dom } g)$, is stronger than $\text{dom } f \cap \text{dom } g \neq \emptyset$ (which is necessary for $f+g$ to be proper) and weaker than $\text{dom } f \cap \text{int}(\text{dom } g) \neq \emptyset$ or $\text{int}(\text{dom } f) \cap \text{dom } g \neq \emptyset$ (which is, in practice, easier to check).

Corollary 8 *Let $f,g : \mathbb{R}^n \rightarrow (-\infty,\infty]$ be ccp with $f+g$ bounded below. If $0 \in int(dom\ f^* + dom\ g^*)$, then (11) is an equality and the infimum of $\inf_y\{f(y)+g(y)\}$ is attained.*

Proof We apply Theorem 7 to $f^*(z)$ and $g^*(-z)$, noting that $\text{dom } g^*(-z) = -\text{dom } g^*(z)$, and that $f+g$ bounded below implies $f^*(z)+g^*(-z)$ is bounded below, by Equation 11. This shows that

$$\inf_z\{f^*(z)+g^*(-z)\} + \inf_y\{f^{**}(y)+g^{**}(y)\} \quad \geq \quad 0$$

is an equality; applying Theorem 5 proves the result. ■

Combining the above two corollaries yields the following variant which we will later apply to learning problems.

Corollary 9 *Let $f,g : \mathbb{R}^n \rightarrow (-\infty,\infty]$ be ccp with $f+g$ bounded below. If $0 \in int(dom\ f - dom\ g)$ or $0 \in int(dom\ f^* + dom\ g^*)$, then*

$$\inf_{y,z}\{f(y)+g(y)+f^*(z)+g^*(-z)\} = 0,$$

and all minimizers y,z satisfy the complementarity equations:

$$\begin{aligned} f(y) - y^t z + f^*(z) &= 0 \\ g(y) + y^t z + g^*(-z) &= 0. \end{aligned}$$

Additionally, if $0 \in int(dom\ f - dom\ g)$ and $0 \in int(dom\ f^* + dom\ g^*)$ then a minimizer (y,z) exists.

We are primarily interested in examples where f and g are ccp, both bounded below, and satisfy both sufficiency conditions. In this case, we see that the minimality conditions of (11) are given by a pair of coupled complementarity equations, each being dependent on only one of the two functions f and g. In the simple case where f and g are both differentiable, these complementary equations are nothing more than $z = \nabla f(y)$ and $-z = \nabla g(y)$, which is clearly the minimality condition for $f(y)+g(y)$. The value of these relations is their generality to non-smooth functions. In our applications to learning, we will be considering the above equations with the regularizer and the loss function taking on the roles of f and g.

3.3 Functions with Auxiliary Parameters

We are frequently interested in functions of the form $h'(y) = \inf_u h(y,u)$. If the infimum is attained for all y where $h'(y)$ is finite, then we say h' is *exact*. We will study the properties of h' through those of h.

Lemma 10 *If* $h : \mathbb{R}^n \times \mathbb{R}^m \to [-\infty, \infty]$ *with* $h'(y) = \inf_u h(y,u)$ *then* $h'^*(z) = h^*(z,0)$.

Proof $h^*(z,0) = \sup_{y,u}\{y^t z - h(y,u)\} = \sup_y\{y^t z - h'(y)\} = h'^*(z)$. ■

It is important to note that $h(y,u)$ being convex in y for fixed u does *not* guarantee that h' is convex. If h is ccp then h' is convex and dom $h' \neq \emptyset$, however, this does not guarantee that h' is exact, closed, or that $h' > -\infty$ (i.e., that h' is proper). We can obtain such guarantees by studying projections of dom h^* onto a subset of its variables.

Lemma 11 *Let* $h : \mathbb{R}^n \times \mathbb{R}^m \to (-\infty, \infty]$ *be ccp with* $h'(y) = \inf_u h(y,u)$. *If* $W = \{w \in \mathbb{R}^m : \exists z \in \mathbb{R}^n, (z,w) \in$ *dom* $h^*\}$ *then* $0 \in W \Rightarrow \forall y, h'(y) > -\infty$.

Proof $\exists y, h'(y) = -\infty \Rightarrow \forall z, h^*(z,0) = \infty \Rightarrow 0 \notin W$. ■

Corollary 12 *Let* h, h' *be as in Lemma 11.*

$$z \in \partial h'(y) \text{ and } h'(y) = h(y,u) \Leftrightarrow (z,0) \in \partial h(y,u).$$

Proof If $h'(y) - y^t z + h'^*(z) = 0$ and $h'(y) = h(y,u)$ then $h(y,u) - y^t z + h^*(z,0) = 0$ and thus $(z,0) \in \partial h(y,u)$. Conversely, if $h(y,u) - y^t z + h^*(z,0) = 0$, since $h(y,u) \geq h'(y)$, we have $0 \geq h'(y) - y^t z + h'^*(z)$. Thus $h'(y) - y^t z + h'^*(z) = 0$ and $h'(y) = h(y,u)$. ■

Lemma 13 *Let* h, h' *be as in Lemma 11. If* $0 \in int(W)$ *then* h' *is ccp and exact.*

Proof For fixed $y \in$ dom h', define $g_y(y',u) = \begin{cases} 0 & y' = y \\ \infty & \text{else} \end{cases}$. It is straightforward to see that $g_y^*(z,w) = \begin{cases} y^t z & w = 0 \\ \infty & \text{else} \end{cases}$, and dom $g_y^* = \mathbb{R}^n \times \{0\}^m$. Hence, dom h^* + dom $g_y^* = \mathbb{R}^n \times W$, and Corollary 8 applies:

$$\begin{aligned} \inf_u h(y,u) &= \inf_{d,u}\{h(d,u) + g_y(d,u)\} \\ &= -\inf_{z,w}\{h^*(z,w) + g_y^*(-z,-w)\} \\ &= -\inf_z\{h^*(z,0) - y^t z\} \\ &= \sup_z\{y^t z - h^*(z,0)\} \\ &= h'^{**}(y), \end{aligned}$$

and there exists d,u which attain $\inf_{d,u}\{h(d,u) + g(d,u)\}$, hence $h(y,u) = h'(y) = h'^{**}(y)$. ■

3.4 Infimal Convolution, Indicator and Support Functions

We introduce the notion of *infimal convolution*, an idea which will play a key role throughout this work.

Definition 14 (infimal convolution) *For $f,g:\mathbb{R}^n \to (-\infty,\infty]$, we define $f \star g:\mathbb{R}^n \to [-\infty,\infty]$, the* infimal convolution *of f and g, by*

$$(f \star g)(y) = \inf_{y'}\{f(y-y')+g(y')\}. \tag{12}$$

We say $f \star g$ is *exact* if the infimum of (12) is attained whenever $(f \star g)(y)$ is finite. If $f \star g$ is exact and $(f \star g)(y)$ is finite, we write $(f \star g)(y) = f(y-y')+g(y')$, implicitly defining y' as a minimizer of (12). The following theorem relates optimality conditions for f and g to optimality conditions for $f \star g$.

Theorem 15 *Let $f,g:\mathbb{R}^n \to (-\infty,\infty]$ be ccp.*

- *$(f \star g)^*(z) = f^*(z)+g^*(z)$. If $0 \in$ dom $f^* -$ dom g^*, then $f \star g > -\infty$.*
- *$z \in \partial(f \star g)(y)$ and $(f \star g)(y) = f(y-y')+g(y') \Leftrightarrow z \in \partial f(y-y') \cap \partial g(y')$.*
- *If $0 \in$ int(dom $f^* -$ dom g^*), then $f \star g = (f^*+g^*)^*$ and is exact (as well as ccp).*

Proof Let $h(y,y') = f(y-y')+g(y')$. $(f \star g)(y) = \inf_{y'} h(y,y')$, hence $f \star g$ is convex. It is straightforward to show that $h^*(z,z') = f^*(z)+g^*(z+z')$. Lastly, define $W = \{z' \in \mathbb{R}^n : (z,z') \in \text{dom } h^*\} = \text{dom } f^* - \text{dom } g^*$. With these results in place, we specialize the previous results.

The first claim is by Lemma 11 and the third by Lemma 13. The second is Corollary 12 with the additional observation:

$$\begin{aligned} h(y,y') - y^t z + h^*(z,0) &= 0 \\ \Leftrightarrow f(y-y')+g(y') - z^t(y-y'+y') + f^*(z)+g^*(z) &= 0 \\ \Leftrightarrow \left[f(y-y') - z^t(y-y') + f^*(z)\right] + \left[g(y') - z^t(y') + g^*(z)\right] &= 0 \end{aligned}$$

Since the two bracketed terms are non-negative (by Theorem 6), the last line is equivalent to $f(y-y') - z^t(y-y') + f^*(z) = g(y') - z^t(y') + g^*(z) = 0$. ■

Many functions of interest can be expressed in terms of infimal convolutions of simpler functions.

There are a number of useful auxiliary functions and sets one may define relative to a given set C.

Definition 16 (indicator functions, support functions, and polarity) *For any non-empty set $C \subset \mathbb{R}^n$, the* indicator function *δ_C, the support function σ_C, and the* polar *of C, C° are given by*

$$\begin{aligned} \delta_C(y) &= \begin{cases} 0 & y \in C \\ \infty & y \notin C \end{cases} \\ \sigma_C(y) &= \sup_{z \in C} z^t y \\ C^\circ &= \{z \in \mathbb{R}^n : \forall y \in C, y^t z \leq 1\}. \end{aligned}$$

Indicator functions allow us to work entirely with unconstrained functions—a problem of the form "minimize $f(y)$ subject to $y \in C$" becomes "minimize $f(y) + \delta_C(y)$." The support function $\sigma_C(y)$ has a simple interpretation as the largest projection of any element of C onto the line generated by y. Indicator functions, support functions, and polars are closely related, as the following lemma shows.

Lemma 17 *Let $C \subset \mathbb{R}^n$ be non-empty. σ_C is ccp, $\delta_C^* = \sigma_C$, C° is closed and convex, and $0 \in C^\circ$.*

Proof C non-empty implies $\sigma_C(0) = 0$, so σ_C is proper. Since epi $\sigma_C = \bigcap_{y \in C}$ epi $H_{y,0}$ and arbitrary intersections of closed and convex sets are closed and convex, σ_C is ccp.

By definition, $\delta_C^*(z) = \sup_y \{y^t z - \delta_C(y)\} = \sup_{y \in C} y^t z = \sigma_C(z)$.

$C^\circ = \{z \in \mathbb{R}^n : \sigma_C(z) \leq 1\}$ is closed and convex, since it is the level set of a closed, convex function, and $\sigma_C(0) = 0$ implies $0 \in C^\circ$. ■

If C is closed, convex and non-empty, then δ_C is ccp, and Theorem 6 applies:

$$\begin{aligned} z \in \partial\delta_C(y) \Leftrightarrow y \in \partial\sigma_C(z) \quad &\Leftrightarrow \quad \delta_C(y) - y^t z + \sigma_C(z) = 0 \\ &\Leftrightarrow \quad y \in C, \forall y' \in C, z^t(y' - y) \leq 0. \end{aligned}$$

If C is a cone then $\sigma_C = \delta_{C^\circ}$, and $C^\circ = \{z \in \mathbb{R}^n : \forall y \in C, z^t y \leq 0\}$. If C is a vector subspace (a special case of a cone), then $C^\circ = C^\perp = \{z \in \mathbb{R}^n : \forall y \in C, z^t y = 0\}$.

Lemma 18 *Let $A, B \subset \mathbb{R}^n$ be closed, convex and non-empty.*

$$\begin{aligned} \delta_{A+B} &= \delta_A \star \delta_B & \sigma_{A+B} &= \sigma_A + \sigma_B \\ \delta_{A\cap B} &= \delta_A + \delta_B & \sigma_{A\cap B} &= \sigma_A \star \sigma_B \quad (\mathit{if}\ 0 \in \mathit{int}(A - B)) \\ \sigma_{A\cup B} &= \sigma_{A\oplus B} & \delta_{A\cup B}^{**} &= \delta_{A\oplus B} \end{aligned}$$

where $A \oplus B$ is the closure of the convex hull of $A \cup B$.

Proof The identities for $\delta_{A\cap B}$, δ_{A+B}, and $\sigma_{A\cup B}$ are easily shown. The others are from conjugation, (with $\sigma_{A\cap B}$ requiring Theorem 15, and hence the sufficiency condition). ■

A number of functions can be built up from support functions. For example, any norm can be defined as the support function of a closed convex set (namely, the unit ball of the associated dual norm). Table 2 contains common examples.

3.5 Summary

Figure 3 provides a concise summary of some of the most important ideas we have presented in this section. In the summary, we ignore necessary technical conditions, but we emphasize that all the applications of convex analysis in this paper refer to and demonstrate the relevant technical conditions. We believe this summary will be useful, especially to readers unfamiliar with the abstract theory of convex functions and conjugacy.

Keep in mind that if f and g are differentiable, much of the theory given here reduces to finding a y such that $\nabla f(y) + \nabla g(y) = 0$. In a very real sense, the entire purpose of the development in this section is to extend intuitions about the unconstrained optimization of differentiable functions to the constrained and non-differentiable setting. This is crucial, as the only unconstrained differentiable examples we are aware of in learning theory are regularized least squares and logistic regression.

C	$\sigma_C(y)$
$\lambda C, \lambda \geq 0$	$\lambda\sigma_C(y)$
AC	$\sigma_C(A^t y)$
$\mathbb{B}_p$	$\|\|y\|\|_{\frac{p}{p-1}}$
$\mathbb{R}^n$	$\delta_{\{0\}}(y)$
$\mathbb{R}^n_{\geq 0}$	$\delta_{\mathbb{R}^n_{\leq 0}}(y)$
$[-1,1]^n$	$\sum_i \|y_i\|$
$[0,1]^n$	$\sum_i (y_i)_+$
K	$\delta_{K^\circ}(y)$

Table 2: Support functions for common convex sets. C and C' are arbitrary non-empty closed convex sets. K is an arbitrary non-empty closed convex cone. A is an arbitrary matrix.

Summary of Key Convex Analysis Concepts

- **(Fenchel-Legendre Conjugate)** $f^*(z) = \sup_y \{y^t z - f(y)\}$.
- **(Biconjugation)** $f^{**} = f$.
- **(Fenchel-Young Theorem)** $f(y) - y^t z + f^*(z) \geq 0$, and $z \in \partial f(y) \Leftrightarrow y \in \partial f^*(z) \Leftrightarrow f(y) - y^t z + f^*(z) = 0$.
- **(Fenchel Duality)** The minimizer of $f(y) + g(y)$ satisfies

$$\begin{aligned} f(y) - y^t z + f^*(z) &= 0 \\ g(y) + y^t z + g^*(-z) &= 0 \end{aligned}$$

  for some z which also minimizes $f^*(z) + g^*(-z)$.
- **(Auxiliary Parameters:** $h'(y) = \inf_u h(y,u)$**)** $h'^*(z) = h^*(z,0)$.
- **(Infimal Convolutions:** $(f \star g)(y) = \inf_{y'} \{f(y-y') + g(y')\}$**)** $(f \star g)^* = f^* + g^*$.
- **(Indicator and Support Functions)** $\delta_C(y) = \begin{cases} 0 & y \in C \\ \infty & y \notin C \end{cases}$, and $\sigma_C(z) = \sup_{y \in C} y^t z$. $\delta_C^* = \sigma_C$. If C is a vector space, $\delta_C^* = \delta_{C^\perp}$.

Figure 3: Summary of key notions of convex conjugacy. In this figure, we assume that all necessary technical conditions (convexity and closedness of functions, exactness of infimal convolutions, etc.) are met; the technical conditions are given in full in the text. All examples given in this paper satisfy the necessary technical conditions. Of course, when considering new examples, the conditions must be checked.

4. Tikhonov Regularization

We now specialize the general theory of Fenchel duality to the inductive learning scenario.

Definition 19 (loss functions and regularization) *A* loss function *is any closed convex function* $v : \mathbb{R} \to (\infty, \infty]$*, which is bounded below and finite at* 0*.*

A regularization *is any closed convex function* $R : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$*, such that* $R(0) = 0$*.*

Typically, a loss function, v_i, represents the penalty for a value mismatching some prescribed value or set of values (e.g., $v_i(y_i) = (y_i - Y_i)^2$), while a regularization represents a measure of *non-smoothness* among a set of values $y_1, \ldots, y_n$.

Clearly, regularizations and loss functions are closed under addition. We can also show that the conjugates of regularizations and loss functions are, respectively, regularizations and loss functions.

Lemma 20 *If v is a loss function, then v^* is a loss function. If R is a regularization, then R^* is a regularization.*

Proof Since v^* and R^* are closed and convex, we need only show that v^* is bounded below and finite at 0 and that $R^*(z) \geq 0$ with $R^*(0) = 0$.

v is bounded below, so $v^*(0) = -\inf_y v(y)$ is finite. $v(0) = -\inf_z v^*(z)$ is finite, so v^* is bounded below.

Since $R(0) = -\inf_z R^*(z) = 0$, R^* is non-negative and $R^*(0) = -\inf_y R(y) = 0$. ■

As a consequence, regularizations and losses are closed under infimal convolution.

The following lemma shows that if the loss function can be decomposed over data points, then its conjugate can likewise be decomposed.

Lemma 21 *Let $V : \mathbb{R}^n \to (-\infty, \infty]$ be given by $V(y) = \sum_{i=1}^n v_i(y_i)$ for loss functions v_i. Then $V^*(z) = \sum_{i=1}^n v_i^*(z_i)$.*

Proof A direct consequence of the definition,

$$\begin{aligned} V^*(z) &= \sup_y \left\{ y^t z - \sum_i v_i(y_i) \right\} \\ &= \sum_i \sup_{y_i} \{ y_i z_i - v_i(y_i) \} \\ &= \sum_i v_i^*(z_i). \end{aligned}$$

■

We are now able to state the main theorem that we will use to study regularization problems.

Theorem 22 (regression Fenchel duality) *Let $R : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ be a regularization and $V(y) = \sum_{i=1}^n v_i(y_i)$ for loss functions $v_i : \mathbb{R} \to \mathbb{R}$. If $0 \in int(dom\ R - dom\ V)$ or $0 \in int(dom\ R^* + dom\ V^*)$ then*

$$\inf_{y,z} \{ R(y) + V(y) + R^*(z) + V^*(-z) \} = 0$$

with all minimizers y, z satisfying the complementarity equations:

$$R(y) - y^t z + R^*(z) = 0 \tag{13}$$

$$v_i(y_i) + y_i z_i + v_i^*(-z_i) = 0. \tag{14}$$

Additionally, if $0 \in int(dom\ R - dom\ V)$ and $0 \in int(dom\ R^* + dom\ V^*)$ then a minimizer exists.

loss	dual loss	optimality condition
$v(y)$	$v^*(-z)$	$v(y)+yz+v^*(-z)=0$
$f(Y-y)$	$f^*(z)-zY$	$f(Y-y)+(Y-y)(-z)+f^*(z)=0$
$f(1-yY)$	$f^*\left(\frac{z}{Y}\right)-\frac{z}{Y}$	$f(1-yY)+(1-yY)\frac{-z}{Y}+f^*\left(\frac{z}{Y}\right)=0$
$\frac{1}{2}y^2$	$\frac{1}{2}z^2$	$y+z=0$
$\lvert y\rvert$	$\delta_{[-1,1]}(z)$	$\lvert y\rvert+yz=0, z\in[-1,1]$
$(y)_+$	$\delta_{[-1,0]}(z)$	$(y)_+ +yz=0, z\in[-1,0]$
$\frac{1}{2}(y)_+^2$	$\delta_{\mathbb{R}_{\leq 0}}(z)+\frac{1}{2}z^2$	$(y)_+ +z=0, z\leq 0$
$(\lvert y\rvert-\varepsilon)_+$	$\delta_{[-1,1]}(z)+\varepsilon\lvert z\rvert$	$\left\{\begin{array}{ll} \lvert y\rvert\geq\varepsilon & z+\operatorname{sign}(y)=0 \\ \lvert y\rvert\leq\varepsilon & z\in[-1,1] \end{array}\right\}$
$\frac{1}{2}(Y-y)^2$	$\frac{1}{2}z^2-zY$	$y+z=Y$
$\lvert Y-y\rvert$	$\delta_{[-1,1]}(z)-zY$	$\lvert Y-y\rvert=(Y-y)z, z\in[-1,1]$
$\lvert 1-yY\rvert$	$\delta_{[-1,1]}\left(\frac{z}{Y}\right)-\frac{z}{Y}$	$\lvert 1-yY\rvert=(1-yY)\frac{z}{Y}, \frac{z}{Y}\in[-1,1]$
$(1-yY)_+$	$\delta_{[0,1]}\left(\frac{z}{Y}\right)-\frac{z}{Y}$	$(1-yY)_+=(1-yY)\frac{z}{Y}, \frac{z}{Y}\in[0,1]$
$\frac{1}{2}(1-yY)_+^2$	$\delta_{\mathbb{R}_{\geq 0}}\left(\frac{z}{Y}\right)+\frac{1}{2}\frac{z^2}{Y^2}-\frac{z}{Y}$	$(1-yY)_+=\frac{z}{Y}, \frac{z}{Y}\geq 0$
$(\lvert Y-y\rvert-\varepsilon)_+$	$\delta_{[-1,1]}(z)+\varepsilon\lvert z\rvert-zY$	$\left\{\begin{array}{ll} \lvert Y-y\rvert\geq\varepsilon & z=\operatorname{sign}(Y-y) \\ \lvert Y-y\rvert\leq\varepsilon & z\in[-1,1] \end{array}\right\}$
$\log(1+\exp(-yY))$	$\delta_{[0,1]}\left(\frac{z}{Y}\right)+\frac{z}{Y}\log\frac{z}{Y}+$ $\left(1-\frac{z}{Y}\right)\log\left(1-\frac{z}{Y}\right)$	$1=(1+\exp(-yY))\frac{z}{Y}$

Table 3: A list of common loss functions and their local optimality conditions. Note that we are considering $v^*(-z)$, not $v^*(z)$. Also note that some of the loss functions can be further simplified under the assumption $Y\in\{-1,1\}$.

We identify the y_i as values taken by the learned function, scored by v_i, and the function values are penalized by $R(y)$. Equation 13 is a complementarity equation in $2n$ variables. Equation 14 is n independent complementarity equations in 2 variables. If R and v_i are differentiable these equations are equivalent to $z=\nabla R(y)$ and $-z_i=\frac{d}{dy}v_i(y_i)$.

4.1 Loss Functions

Table 3 recapitulates many of the loss functions seen in practice. In this paper, we deal exclusively with pointwise losses, so Table 3 is in terms of a single regression value y and a single training value Y with subscripts omitted.

The derivations in Table 3 can all be obtained without much effort from identities and definitions previously stated and earlier derivations in the table. It is helpful to observe that many loss functions can be expressed in terms of infimal convolutions of simple functions. For example, $(y)_+=\sigma_{[0,1]}(y)=(|\cdot|\star\delta_{\mathbb{R}_{\leq 0}})(y)$ and $(|y|-\varepsilon)_+=(|\cdot|\star\delta_{[-\varepsilon,\varepsilon]})(y)$.

It should be noted that these derivations can be checked graphically, since the losses are functions of a single variable. For example, Figure 4 shows a graph of $(1-y)_+$, the well-known *hinge loss*, with supporting hyperplanes. From the figure, it is plain that the only supporting hyperplanes are those with slopes in the interval $[-1,0]$, thus dom $f^*=[-1,0]$, and that $f^*(y)=y$ when $y\in[-1,0]$.

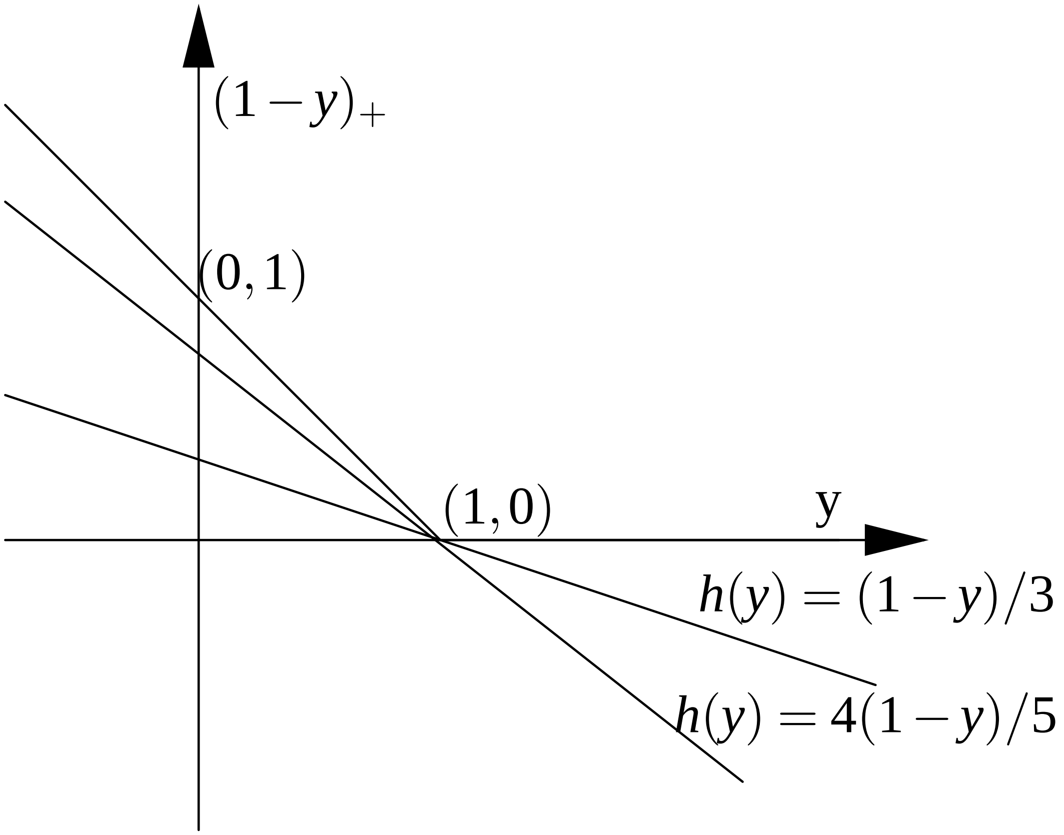

Figure 4: The graph of $(1-y)_+$ with a couple supporting hyperplanes.

4.2 Regularization Functions

By Lemma 20, if $R_1, R_2 : \mathbb{R}^n \to [0,\infty]$ are regularizations, then $R_1 + R_2$ and $R_1 \star R_2$ are, which gives two ways of building more complicated regularizations out of simpler ones. The basic building blocks of regularizations are indicator functions, support functions, and quadratic forms. The identities to keep in mind are $\sigma_C^* = \delta_C$, $Q_A^* = Q_{A^{-1}}$ for non-singular A, and $R_1 \star R_2 = (R_1^* + R_2^*)^*$ (with appropriate conditions from Theorem 15).

We might think of sums of regularizations as adding further penalties on the regression values, as $R \leq R + R'$. On the other hand, infimal convolutions add slack, $R \geq R \star R'$. Since these two operations are conjugates of each other, it is clear that any additional restriction in the primal results in extra freedom in the dual and vice versa.

5. Regularization in Reproducing Kernel Hilbert Spaces

We now turn to our primary example, Tikhonov regularization in a reproducing kernel Hilbert space (RKHS).

reg.	dual reg.	optimality condition
$R(y)$	$R^*(z)$	$R(y) - yz + R^*(-z) = 0$
$R_1 \star R_2(y)$	$R_1^*(z) + R_2^*(z)$	$R_1(y-y') - (y-y')^t z + R_1^*(z) = R_2(y') - y'^t z + R_2^*(z) = 0$
$R(A^t y)$	$R^*(A^\dagger z) + \delta_{A\mathbb{R}^n}(z)$	$R(A^t y) - y^t z + R(A^\dagger z) = 0, z \in A\mathbb{R}^n$
$\sigma_C(y)$	$\delta_C(z)$	$\sigma_C(z) = y^t z, z \in C$
$\sigma_C(A^t y)$	$\delta_{AC}(z)$	$\sigma_C(z) = y^t z, z \in C$
$Q_A(y) = \frac{1}{2} y^t A y$	$Q_{A^\dagger}(z) + \delta_{A\mathbb{R}^n}(z)$	$z = Ay$
$\|\|y\|\|_p$	$\delta_{\|\|z\|\|_q \leq 1}(z)$	$\|\|y\|\|_p = y^t z, \|\|z\|\|_q \leq 1, (\frac{1}{p} + \frac{1}{q} = 1)$
$\|\|A^t y\|\|_1$	$\delta_{A[-1,1]^n}(z)$	$\|\|A^t y\|\|_1 = y^t z, z \in A[-1,1]^n$
$\frac{1}{2}\|\|y\|\|_p^2$	$\frac{1}{2}\|\|z\|\|_q^2$	$\|\|y\|\|_p^2 = \|\|z\|\|_q^2 = y^t z, (\frac{1}{p} + \frac{1}{q} = 1)$

Table 4: Common regularization choices.

5.1 Optimality Conditions for RKHS Regularization

Recall that the "standard" approach in machine learning is to start with a Tikhonov regularization problem over an RKHS (2), to invoke the representer theorem to show that the solution can be expressed as a collection of coefficients (3), and to write optimization problems in terms of these coefficients. We start with a mathematical program in terms of the y_i, using a regularization $Q_{K^{-1}}$ (with K symmetric, positive definite); in this framework, we are able to state general optimality conditions and derive a very clean form of the representer theorem.

Specialized to the RKHS case, we are considering primal problems of the form

$$\inf_{y \in \mathbb{R}^n} \left\{ \frac{1}{2} \lambda y^t K^{-1} y + V(y) \right\},$$

where we have made the target labels Y_i implicit in V. The associated dual problem is

$$\inf_{z \in \mathbb{R}^n} \left\{ \frac{1}{2} \lambda^{-1} z^t K z + V^*(z) \right\}.$$

Equation 13 specializes to

$$\begin{aligned} \frac{1}{2}\lambda y^t K^{-1} y - y^t z + \frac{1}{2}\lambda^{-1} z^t K z &= 0 \\ \frac{1}{2}(y - \lambda^{-1} K z)^t (\lambda K^{-1} y - z) &= 0. \end{aligned}$$

and thus we obtain the optimality conditions $z = \lambda K^{-1} y$ (or $y = \lambda^{-1} K z$). This demonstrates that whenever we are performing regularization in an RKHS, the optimal y values at the training points can be obtained by multiplying the optimal dual variables z by $\lambda^{-1} K$, *for any convex loss function.*

5.2 RKHS Regularization Examples

In RKHS regularization, we have a regularizer $R(y) = \lambda Q_{K^{-1}}(y)$, with conjugate $R^*(z) = \lambda^{-1} Q_K(z)$, and optimality conditions $y = \lambda^{-1} K z$. By incorporating a loss function, we can derive well-known kernel machines.

5.2.1 REGULARIZED LEAST SQUARES

The *square* loss is given by $v(y_i) = \frac{1}{2}(Y_i - y_i)^2$. Although it is listed in Table 3, we derive the dual loss here, for pedagogical purposes:

$$v^*(-z) = \sup_{y'}\{-y'z - \frac{1}{2}(y' - Y)^2\}.$$

We see that $v^*(-z)$ is quadratic in y', and we can find the sup by taking the derivative:

$$\frac{\partial v^*}{\partial y'} = -z + (Y - y').$$

Setting the derivative to 0, we obtain $y = Y - z$, and substituting, we see that

$$v^*(-z) = -zY + \frac{1}{2}z^2.$$

The optimality equation $v(y_i) + y_i z_i + v^*(z_i)$ reduces to $y + z = Y$, so (13) and (14) become

$$\begin{aligned} y_i + z_i &= Y_i \\ \lambda^{-1}Kz &= y. \end{aligned}$$

Combined,

$$y + \lambda K^{-1}y = \lambda^{-1}Kz + z = Y,$$

leading to the standard regularized least squares formulation.

5.2.2 UNBIASED SVM

The support vector machine arises from the combination of RKHS regularization and the hinge loss $v(y) = (1 - yY)_+$. The dual of the hinge loss, $v^*(-z) = \delta_{[0,1]}(\frac{z}{Y}) - \frac{z}{Y}$, is most easily derived using a graphical approach (see Figure 4), but it can also be derived directly. Regularizing in an RKHS, we have the primal and dual problems:

$$\inf_{y \in \mathbb{R}^n} \left\{ \frac{1}{2}\lambda y^t K^{-1} y + \sum_i (1 - y_i Y_i)_+ \right\}$$
$$\inf_{z \in \mathbb{R}^n} \left\{ \frac{1}{2}\lambda^{-1} z^t K z + \sum_i \left(\delta_{[0,1]}\left(\frac{z_i}{Y_i}\right) - \frac{z_i}{Y_i} \right) \right\}.$$

We see how the δ function in the dual loss provides the well-known “box constraints” in the SVM dual.

Given $v(y), v^*(-z)$, it is also straightforward to derive the optimality condition

$$(1 - yY)_+ = (1 - yY)\frac{z}{Y} \quad \text{and} \quad \frac{z}{Y} \in [0,1].$$

A case analysis on the sign of $(1-yY)$ yields

$$\begin{aligned}
(1-y_iY_i) < 0 &\iff \frac{z_i}{Y_i} = 0 \\
(1-y_iY_i) = 0 &\iff \frac{z_i}{Y_i} \in [0,1] \\
(1-y_iY_i) > 0 &\iff \frac{z_i}{Y_i} = 1.
\end{aligned}$$

With the regularization condition $y = \lambda^{-1}Kz$, we have derived the KKT conditions for the SVM. To obtain a more "standard" formulation, we could introduce variables $\alpha_i \equiv \frac{z_i}{Y_i}$, $c = \lambda^{-1}z$, eliminating y from the system (in favor of z and α), and introducing a "regularization constant" $C = \frac{1}{2\lambda}$.

In this section, we have derived an "unbiased" SVM. In practice, the SVM usually includes an unregularized bias term b. We show how to handle the bias term b in Section 9.1.

6. L1 Regularizations

Tikhonov regularization in a reproducing kernel Hilbert space is the most common form of regularization in practice, and has a number of mathematical properties which make it especially nice. In particular, the representer theorem makes RKHS regularization essentially the only case where we can start with a problem of optimizing an infinite-dimensional function in a function space, and obtain a finite-dimensional representation.[5] Nevertheless, other regularizations may be of interest. In these cases, we are generally forced to assume *a priori* that the functions we are looking for are coefficients in some finite dimensional space.

6.1 1-norm Support Vector Machines

As a consequence of the use of the hinge loss, support vector machines yield a sparse solution—points which live outside the margin have $c = z = 0$. Nevertheless, because all points which are errors or live inside the margin are support vectors, in noisy problems, one generally finds that at least a constant fraction of the examples (lower-bounded by the Bayes error rate) are support vectors. Instead of using RKHS regularization, we can a priori choose a finite-dimensional function space of expansion coefficients of kernel functions around the training points $y = \sum_i K(x,x_i)c_i$, and regularize $||c||_1$. If we combine this idea with the hinge loss, we obtain the 1-norm support vector machine (Zhu et al., 2003).

We consider $R(y) = ||K^{-1}y||_1$. If we use the hinge loss (or any other piecewise linear loss function), the resulting mathematical optimization problem is a linear program. The dual regularizer is $R^*(z) = \delta_{K^{-1}[-1,1]^n}(z)$. As usual, we can combine primal and dual regularizers with primal and dual loss functions. Note that in these formulations, the formula $c = \lambda^{-1}z$ *does not hold*: if we solved a problem in y and z, we would need to then solve $c = K^{-1}y$ to obtain c. In practice, we expect that algorithms based directly on c are probably the most useful.

[5] Kernel-based algorithms are made practical by two factors: the representer theorem for RKHS's and kernels which make kernel products (and therefore the representations) easy to compute. One can construct RKHS's, for example, via embeddings, where learning is not practical because of a computationally difficult feature map and inner product. We thank a reviewer for pointing out this distinction.

6.2 Sparse Approximation and Support Vector Regression

In Girosi (1998), an "equivalence" between a certain sparse approximation problem and a modified form of SVM regression is developed. In slightly simplified form, the equivalence is quite easy to derive.

We begin by considering the sparse approximation problem. We are given *noiseless* samples $\{(X_1,Y_1),\dots,(X_n,Y_n)\}$ from a target function f^t which is assumed to live in an RKHS $\mathcal{H}$. We will find a function which is a coefficient expansion around the data points: $f(x) = \sum_i K(x,x_i), c_i$. The goal is to minimize the distance between f^t and $f(x)$ in the RKHS norm,[6] while maintaining sparsity of the coefficient representation c:

$$\inf_{c\in\mathbb{R}^n}\left\{\frac{1}{2}||f^t - f||^2_{\mathcal{H}} + \varepsilon||c||_1\right\} = \frac{1}{2}||f^t||^2_{\mathcal{H}} + \inf_{c\in\mathbb{R}^n}\left\{\frac{1}{2}c^tKc - Y^tc + \varepsilon||c||_1\right\}.$$

where we used basic properties of RKHS and the assumption that $f^t(x) \in \mathcal{H}$ to expand $||f^t - f||^2_{\mathcal{H}}$.

Now we turn to a variant of support vector machine regression. A standard support vector regression machine is obtained when we use a loss $v(y_i) = (|Y_i - y_i| - \varepsilon)_+$, linearly penalizing points which lie outside a tube. Girosi instead considers a "hard margin" variant, where points are *required* to lie inside the tube. His primal problem is:

$$\inf_{y\in\mathbb{R}^n}\left\{\lambda Q_{K^{-1}}(y) + \delta_{[-\varepsilon,\varepsilon]^n}(Y - y)\right\}.$$

The dual to the "hard loss" $v(y_i) = \delta_{[-\varepsilon,\varepsilon]}(Y_i - y_i)$ is given by

$$\begin{aligned} v^*(-z_i) &= \sup_{y'_i}\{-y'_i z_i - \delta_{[Y_i-\varepsilon,Y_i+\varepsilon]}(y'_i)\} \\ &= \max\{-(Y_i-\varepsilon)z_i, -(Y_i+\varepsilon)z_i\} \\ &= -Y_i z_i + \varepsilon|z_i|. \end{aligned}$$

Therefore, the dual problem is

$$\inf_{z\in\mathbb{R}^n}\left\{\lambda^{-1}Q_K(z) - Y^tz + \varepsilon||z||_1\right\} = \lambda\cdot\inf_{z\in\mathbb{R}^n}\left\{Q_K(\lambda^{-1}z) - Y^t(\lambda^{-1}z) + \varepsilon||\lambda^{-1}z||_1\right\}.$$

We see that the sparse approximation variant and the support vector regression variant have optimal values that differ by a constant $\frac{1}{2}||f^t||^2_{\mathcal{H}}$ and a positive multiplier λ, with $c = \lambda^{-1}z$. The sparsity control parameter ε becomes the width of the allowed tube in the regression problem. This makes sense: as we increase ε, we get more sparsity, but our predicted y values must become less tightly constrained in order to achieve that sparsity.[7]

[6] In earlier treatments of sparse decomposition (such as Chen et al., 1995) the goal was to minimize functional distance in L_2 rather than $\mathcal{H}$; this distance (of course) needed to be approximated empirically.

[7] Girosi considered biased regularizations with a free constant b. In order to incorporate this, he was forced to make additional assumptions on f^t and K. While these assumptions gave a formal equivalence between the sparse problem and the biased SVM regression quadratic program, we feel that the simplified version presented here illuminates the essential equivalence much more clearly.

7. Representer Theorem

We can generalize beyond the training points to talk about the predicted values at future points—our own version of the representer theorem. We consider a scenario where we have n training points and m additional unlabelled points and show how the y values at the unlabelled points can be determined via kernel expansions around the training points.

We begin by relating arbitrary regularizations in higher and lower-dimensional spaces.

Lemma 23 *Let $R : \mathbb{R}^{n+m} \to (-\infty, \infty]$ be a regularization, and let $R' = \inf_{y_2 \in \mathbb{R}^m} R(y_1, y_2)$. If $0 \in int(dom\ R^*)$, then R' is ccp and exact, R' is a regularization, and $R'^*(z_1) = R^*(z_1, 0)$. Additionally,*

$$y_1 \in \partial R'^*(z_1) \text{ and } R'(y_1) = R(y_1, y_2) \Leftrightarrow (y_1, y_2) \in \partial R^*(z_1, 0).$$

Proof By Lemmas 11 and 13, we have R' being ccp and exact. Clearly, $\inf_{y_1} R'(y_1) = R'(0) = 0$. The last claim comes from Corollary 12 and Theorem 6. ■

We specialize the lemma to the quadratic form regularizers associated with RKHS:

Corollary 24 *Let $K = \begin{pmatrix} K_{NN} & K_{NM} \\ K_{MN} & K_{MM} \end{pmatrix} \in \mathbb{R}^{(n+m)\times(n+m)}$ be symmetric positive definite where $K_{NN} \in \mathbb{R}^{n\times n}$. For all $y_1 \in \mathbb{R}^n$,*

$$\inf_{y_2 \in \mathbb{R}^m} Q_{K^{-1}}(y_1, y_2) = Q_{K_{NN}^{-1}}(y_1), \tag{15}$$

where the infimum is (uniquely) attained at $y_2 = K_{MN} K_{NN}^{-1} y_1$.

Proof Let $R(y_1, y_2) = Q_{K^{-1}}(y_1, y_2)$, and $R'(y_1) = \inf_{y_2} R(y_1, y_2)$. $R^*(z_1, z_2) = Q_K(z_1, z_2)$, thus dom R^* $= \mathbb{R}^{n+m}$ and R' is therefore closed and exact. $R'^*(z_1, 0) = Q_K(z_1, 0) = Q_{K_{NN}}(z_1)$, so $R'(y_1) = Q_{K_{NN}^{-1}}(y_1)$, hence (15). Given y_1, let z_1 satisfy $K_{NN} z_1 = y_1$. Then, $y_1 \in \partial R'^*(z_1)$. Also, $\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \in \partial R^*(z_1, 0)$ iff $y_2 = K_{MN} z_1$. By the last part of Lemma 23, $y_2 \in \text{argmin}_{y_2} Q_{K^{-1}}(y_1, y_2)$. Since z_1 is unique given y_1, y_2 is unique. ■

We are now able to prove an optimization-centered variant of the representer theorem.

Theorem 25 *Let $X_1, \ldots, X_n, \ldots, X_{n+m} \in \mathbb{R}^d$ with $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ an symmetric positive definite kernel function. Let $K \in \mathbb{R}^{(n+m)\times(n+m)}$ be defined by $K_{ij} = k(X_i, X_j)$ for $1 \le i, j \le n+m$ with $K_{NN} \in \mathbb{R}^{n\times n}$ the upper left submatrix of K.*

For any function $V : \mathbb{R}^n \to (-\infty, \infty]$,

$$\inf_{y \in \mathbb{R}^{n+m}} \{Q_{K^{-1}}(y) + V(y_1, \ldots, y_n)\} = \inf_{y' \in \mathbb{R}^n} \left\{Q_{K_{NN}^{-1}}(y') + V(y')\right\},$$

where $y_i = y'_i$ for $1 \le i \le n$ and $y_i = \sum_{j=1}^n k(X_i, X_j) c_j$ with $c = K_{NN}^{-1} y'$.

Proof Since the loss term V is independent of the last m components of y,

$$\begin{aligned} \inf_{y \in \mathbb{R}^{n+m}} \{Q_{K^{-1}}(y) + V(y_1, \ldots, y_n)\} &= \inf_{y' \in \mathbb{R}^n} \left\{ \inf_{y'' \in \mathbb{R}^m} Q_{K^{-1}}(y', y'') + V(y') \right\} \\ &= \inf_{y' \in \mathbb{R}^n} \left\{ \inf_{y'' \in \mathbb{R}^m} Q_{K_{NN}^{-1}}(y') + V(y') \right\}, \end{aligned}$$

by Corollary 24. ■

Our variant of the representer theorem shows that any point y_i which appears in the regularization but not the loss term can be determined from an expansion of the kernel function about the training points. Thus, solving the n variable optimization problem simultaneously "solves" all $n+m$ optimization problems with any additional set of m lossless points. This was made possible by the "subsetting" property of quadratic form regularizers, developed in Corollary 24.

An alternate way of thinking about this version of the representer theorem is that for regularization in an RKHS, the inductive and transductive cases "match": if we are given n training points and m testing points in advance, we can either train the standard n point optimization problem and use the coefficients c to compute the y values at the test points, or we can directly solve a larger optimization problem in which the test points appear in the regularization but not in the loss.

In RKHS regularization, we define $c \equiv K^{-1}y = \lambda^{-1}z$; this is justified by the representer theorem. We see that not only can we compute the values at test points, but the coefficients c have a direct and simple relation to the dual variables z. As we will see in later sections, when we explore other regularizations and kernel approximations, when we leave the RKHS setting, the simple relation between c and z is severed.

It is worthwhile to put our result in the context of more "standard" forms of the representer theorem. In its most general form (see Schölkopf et al. (2001) or Belkin et al. (2004) for proofs of this statement with a slightly different notation and emphasis), the representer theorem states that the optimal solution to an optimization problem of the form:

$$\min_{f \in \mathcal{H}} \left\{ \lambda ||f||_K^2 + F(f(X_1), \ldots, f(X_n)) \right\}, \tag{16}$$

where F is an arbitrary functional that depends only on the values f takes at the X_i, will have a solution of the form

$$f(x) = \sum_{i=1}^{n} c_i K(x, X_i), \tag{17}$$

and that the infinite-dimensional problem of finding an optimal $f \in \mathcal{H}$ can be reduced to the finite-dimensional problem of finding the optimal c_i. Traditionally, F is taken to be a loss function over labelled training points, but this is not required—for example, if the training set contains a mix of labelled and unlabelled points, F may be the sum of a loss function over the labelled points and some additional regularizer over the unlabelled points (see Belkin et al. (2004); Rahimi et al. (2005), and also Section 7.1; note that in Theorem 25 we define V as arbitrary and do not require it to be a loss function). The proof essentially proceeds by defining f to be a sum of kernel expansions (Equation 17) around the training points and a "remainder" which is orthogonal to these kernel expansions, and then showing that the orthogonal remainder must be the zero function. In our development, it is clear that solving

$$\min_{y \in \mathbb{R}^n} \left\{ y^t K^{-1} y + V(y) \right\}, \tag{18}$$

and defining $y = Kc$ will find the optimal solution to (16) of the form (17). Any function in the RKHS can be represented as a possibly infinite expansion of the form (17); Theorem 25 shows that if we add any (finite) set of additional points to the representation, none of the predicted training values change, and the expansion coefficients at the new points vanish. A simple limiting argument allows us to extend to infinite sets of additional points, demonstrating that the solution to (16) can be found by solving (18).

7.1 Semi-supervised and Transductive Learning

In this section we discuss in detail the somewhat subtle relationships between semi-supervised, transductive and inductive learning, quadratic regularizers, and the representer theorem.

In standard usage, inductive and semi-supervised learning are viewed as problems of taking a training set (fully labelled for the inductive case, partially unlabelled for the semi-supervised case) and learning a *function* that classifies new examples. In contrast, transductive learning is often viewed as the problem of predicting the labels at unlabelled points, given the locations of those points *in advance*. It must be understood that this distinction is somewhat artificial, based on notions of what we consider a function. In particular, given any transductive algorithm $\mathcal{A}$, we obtain a semi-supervised algorithm for classifying new points by rerunning $\mathcal{A}$ with all points seen so far whenever a prediction is needed. From this bird's-eye view, all problems about "out of sample" extensions for transductive algorithms (see for example Bengio et al. (2003)) vanish. This is a perfectly legitimate semi-supervised algorithm, in that it provides a *function* for computing the values at new points; however, because evaluating the function requires solving an optimization problem, many people are uncomfortable with this algorithm. At the present time, it seems that it is common usage in the machine learning community to accept predictive functions which are weighted sums of expansions around a fixed basis as legitimate semi-supervised solutions, but to reject predictive functions which require solving optimization problems as being merely transductive algorithms. This viewpoint is in some sense arbitrary, but also reflects genuine computational concerns; for the remainder of this section, we refer to transductive and inductive algorithms as the terms are commonly used.

Given an arbitrary transductive algorithm, it is possible to turn it into an inductive algorithm by embedding the optimization in an RKHS. An excellent example is manifold regularization Belkin and Niyogi (2003). As originally formulated, manifold regularization is defined only for points on a graph, and is therefore a transductive algorithm. In Belkin et al. (2004), the authors construct an inductive algorithm by considering the optimization problem:

$$\min_{y \in \mathbb{R}^{m+n}} \left\{ \lambda_1 y^t K^{-1} y + \lambda_2 y^t L y + V(y_N) \right\}, \tag{19}$$

where we have n labelled and m unlabelled points, y_N are the predicted values at (just) the labelled points, and L is the graph Laplacian of (all) the data. The term $y^t L y$ is an additional smoothness penalty on the predicted values that respects the notion that we expect the data to live on or near a low-dimensional manifold. By the standard representer theorem, solving the above optimization is equivalent to finding a function with minimal sum of its RKHS norm and the "loss function" $\lambda_2 y^t L y + V(y_N)$. Similar extensions of straightforward RKHS regularizers with additional penalties are known for time series (Rahimi et al., 2005) and structured prediction (Altun et al., 2005).[8] In fact, whenever the additional regularization term is a quadratic form Q_L, we can view the semi-supervised algorithm as finding the optimal function in a data-dependent "warped" RKHS (see Sindhwani et al. (2005) for details).

While we certainly agree that Problem 19 gives a finite-dimensional solution to an infinite-dimensional problem, we question whether this problem is the "right" problem to solve. In particular, we consider two "strategies" for classifying new points:

[8] We have not seen many examples in the machine learning literature of researchers treating the outputs y as the predictive values. The above algorithms are generally phrased in terms of c, or sometimes in terms of $f(x)$, with $f(x)$ viewed as dependent on c.

- **(Inductive)** Solve Problem 19. Compute c via $y = Kc$, yielding a "function" $f(x) = \sum_i c_i K(x, x_i)$. Use f to predict values at new points.

- **(Transductive)** When presented with a new point, compute an augmented form of optimization Problem 19 by adding the new point to the RKHS and additional regularization terms. Predict the value at the new point as the associated optimal value in the augmented optimization problem.

Given sufficient computational resources, we consider the transductive approach to be the "natural" choice, in that it treats previous and future unlabelled points on an equal basis. The inductive algorithm may be appropriate as a computationally more tractable substitute, but it is crucial to keep in mind that we do not expect the inductive and transductive approaches to give the same answer. This is in direct contrast to the case of standard RKHS supervised learning, where Theorem 25 shows that the inductive and transductive approaches *give the same answer*. For RKHS supervised learning, unlabelled points do not contribute to the loss (we implicitly assume $v_i(y_i) = 0$ for unlabelled points). On the other hand, when we consider a data-dependent smoothness functional over unlabelled data, additional points give us additional information about this smoothness penalizer. In the transductive semi-supervised case, adding the unlabelled point to the optimization problem can change the predictions at previous points, both labelled and unlabelled. We will see another example of this phenomenon in Section 8.

An alternate perspective on problems like Problem 19 is to think of the smoothness penalty $Q_L(y)$ as part of the regularizer, rather than part of the loss. From this viewpoint, we no longer have a (data independent) RKHS regularizer, and so Theorem 25 does not apply.

8. Learning the Kernel

Standard RKHS regularization involves a fixed kernel function k and kernel matrix K. Recently, there has been substantial interest in scenarios in which a kernel matrix or kernel function is learned from the data, simultaneously with learning a prediction function. We first develop some very general results on this case, and then show how these results can easily be used to obtain and generalize results from the recent literature.

We consider a variant of Tikhonov regularization where the regularization term itself contains parameters $u \in \mathbb{R}^p$, which are optimized; we learn the regularization in tandem with the regression. The primal becomes

$$\inf_{y,u} \left\{ R(y,u) + \sum_i v_i(y) \right\}.$$

where $R(y,u)$ is convex in y. Given the independence of the loss from u, we can move the u-infimum inside

$$\inf_{y} \left\{ \inf_{u} R(y,u) + \sum_i v_i(y) \right\}$$

obtaining a new regularization $R'(y) = \inf_u R(y,u)$. We will assume through the remainder that $R(y,u)$ is ccp.

By lemma 11, $R'^*(z) = R^*(z,0)$. If R' is exact then the optimality condition for the regularization becomes

$$R(y,u) - y^t z + R^*(z,0) = 0,$$

or, equivalently, $(z,0) \in \partial R(y,u)$.

We now specialize to the RKHS case, and allow the entire kernel to play the role of the auxiliary variables: $R(y,K) = \lambda Q_{K^{-1}}(y) + F(K)$ for some closed convex F whose domain $\mathcal{K}$ is contained in the $n \times n$ symmetric positive semidefinite matrices. Let $A \bullet B \equiv \sum_{i,j} A_{ij} B_{ij}$ and recall $zz^t \bullet K = z^t K z$. R is ccp and $R^*(z,W) = F^*\left(W + \frac{1}{2}\lambda^{-1} zz^t\right)$ which we can verify by Theorem 5, taking the conjugate

$$\begin{aligned}
R^*(z,W) &= \sup_{y,K} \left\{ y^t z + K \bullet W - \lambda Q_{K^{-1}}(y) - F(K) \right\} \\
&= \sup_{K} \left\{ \sup_{y} \{ y^t z - \lambda Q_{K^{-1}}(y) \} + K \bullet W - F(K) \right\} \\
&= \sup_{K} \left\{ \lambda^{-1} Q_K(z) + K \bullet W - F(K) \right\} \\
&= \sup_{K} \left\{ K \bullet \left(W + \frac{1}{2}\lambda^{-1} zz^t \right) - F(K) \right\} \\
&= F^*\left(W + \frac{1}{2}\lambda^{-1} zz^t \right)
\end{aligned}$$

and biconjugate,

$$\begin{aligned}
\sup_{z,W} \left\{ y^t z + K \bullet W - F^*\left(W + \frac{1}{2}\lambda^{-1} zz^t \right) \right\} &= \\
\sup_{z} \left\{ y^t z + \sup_{W} \left\{ K \bullet W - F^*\left(W + \frac{1}{2}\lambda^{-1} zz^t \right) \right\} \right\} &= \\
\sup_{z} \left\{ y^t z - K \bullet \frac{1}{2}\lambda^{-1} zz^t \right\} + F^{**}(K) &= \\
\sup_{z} \left\{ y^t z - \lambda^{-1} Q_K(z) \right\} + F^{**}(K) &= \\
\lambda Q_{K^{-1}}(y) + F(K) &= R(y,K).
\end{aligned}$$

By Lemmas 11 and 13, $R'^* = F^*\left(\frac{1}{2}\lambda^{-1} zz^t\right)$ and R' is closed and exact if $\forall z \in \mathbb{R}^n, zz^t \in \operatorname{int}(\operatorname{dom} F^*)$. If R' is exact, the condition for optimality for the regularization (i.e., $z \in \partial R'(y)$) is

$$\frac{1}{2}\lambda y^t K^{-1} y + F(K) - y^t z + F^*\left(\frac{1}{2}\lambda^{-1} zz^t\right) = 0$$

$$\left[\frac{1}{2}\lambda y^t K^{-1} y - y^t z + \frac{1}{2}\lambda^{-1} z^t K z\right] + \left[F(K) - \operatorname{tr}\left(\frac{1}{2}\lambda zz^t K\right) + F^*\left(\frac{1}{2}\lambda^{-1} zz^t\right)\right] = 0.$$

Each of the two bracketed terms is non-negative, by Theorem 6; therefore they must both vanish. Hence, $y = \lambda^{-1} K z$ and $\frac{1}{2}\lambda^{-1} zz^t \in \partial F(K)$, equivalently, $K \in \partial F^*\left(\frac{1}{2}\lambda^{-1} zz^t\right)$.

Given an RKHS regularizer, learning K and y simultaneously will be a convex problem as long as F is closed and convex. However, in order to obtain a useful learning algorithm, the function

F must provide a meaningful constraint on allowable kernels, and there must be a mechanism for predicting values at new points. Abusing notation, let F_N and F_M be the versions of F that operate on $n \times n$ and $m \times m$ matrices, respectively (we usually suppress this notation). Suppose that, whenever $m > n$, we have the property that $F_N(A) = \inf_{K:K_{NN}=A} F_M(K)$. In this case, it is straightforward to see that adding additional "testing" points to the regularizer, but not the loss, will not change the objective value, nor will it change the y values at the "training" points. This leads to a transductive algorithm. Furthermore, if the minimizer in the inf is unique, we can use K_N to determine K_M given the new x points, we will have a representer theorem, and the inductive and transductive cases will be identical.

Lanckriet et al. (2004) consider a transductive scenario, where the training and testing points are known in advance. They start with a finite set of kernels $K_1, \ldots, K_k$ (the K_i are over the training and testing sets) and take $F(K) = \delta_{K_{u,c}}(K)$, where

$$K_{u,c} \equiv \left\{ K : K = \sum_i u_i K_i, \operatorname{tr}(K) = c,\ K \succeq 0 \right\}.$$

(Lanckriet et al. (2004) frequently consider $K_{u^+,c}$ as well, where the u are constrained to be nonnegative.) A primary concern in Lanckriet et al. (2004) is showing that when the SVM hinge loss is used, the resulting optimizations can be phrased as semidefinite programming problems. They recognize that it is important to "entangle" the training and testing kernel matrices—if we simply allowed K to range over the entire semidefinite cone, for example, we would obtain kernel matrices which fit the training data well and ignored the testing data. Using a finite combination of kernel matrices essentially means that we are learning a kernel function parametrized by u, and applying this function to both training and testing points. Because Lanckriet et al. (2004) constrain the trace of the *entire* kernel matrix K, adding additional testing points can change the value at training points, and they cannot easily and directly perform induction. If they had instead chosen to constrain the trace of the kernel matrix over only the training points, or simply to constrain the sum of the u_i, they could have obtained a representer theorem, and there would have been agreement between their transductive algorithm and the obvious inductive algorithm. Lanckriet et al. (2004) focus primarily on the SVM hinge loss. In fact, any convex loss can be used, and the resulting optimization problem can be cast as a semidefinite programming problem (Recht, 2006).

While Lanckriet et al. (2004) consider a finitely generated set of kernel matrices, Argyriou et al. (2005) works with a convex set generated by an infinite, continuously parametrized set of kernel functions, the primary example being Gaussian kernels with all bandwidths in some closed interval. Argyriou et al. (2005) show that the optimal kernel function will have a "representation" in terms of $n+1$ basic kernels. Because we have separated our concerns with the kernel function appearing only in the regularization and not in the loss, we are able to give a very simple proof of their result. Given a subset, $\mathcal{K}$, of the $n \times n$ symmetric positive semidefinite matrices, we consider $F(K) = \delta_{\mathcal{K}^\oplus}(K)$, where $\mathcal{K}^\oplus$ is the convex hull of $\mathcal{K}$.

Lemma 26 *Let $z \in \mathbb{R}^n$ and $K \in \mathcal{K}^\oplus$. There exists $\tilde{K} = \sum_{i=1}^{n+1} t_i K_i$ with $t_i \geq 0$ and $\sum_i t_i = 1$ such that $\tilde{K}z = Kz$.*

Proof Let $Y = \{K'z : K' \in \mathcal{K}\}$ and $Y^\oplus = \{K'z : K' \in \mathcal{K}^\oplus\}$. Clearly, $Y^\oplus$ is the convex hull of Y and $Kz \in Y^\oplus$. By Carathéodory's theorem (Rockafellar and Wets (2004), Theorem 2.29), $\exists t \in \mathbb{R}^{n+1}, t_i \geq 0, \sum_i t_i = 1$ such that $Kz = \sum_i t_i y_i$ with $y_i \in Y$ and hence $y_i = K_i z$ for some $K_i \in \mathcal{K}$. ■

$F^* = \sigma_{\mathcal{K}^\oplus}$. Let us assume that $\mathcal{K}^\oplus$ is closed (compactness of $\mathcal{K}$ is sufficient but not necessary). Under this assumption, F is ccp, and the optimality condition from the regularization term becomes

$$y = \lambda^{-1}Kz \quad \text{and} \quad z^tKz = \sigma_{\mathcal{K}^\oplus}(zz^t), K \in \mathcal{K}^\oplus. \tag{20}$$

Corollary 27 *Let $z \in \mathbb{R}^n, K \in \mathcal{K}^\oplus$ and $\tilde{K} \in \mathcal{K}^\oplus$ be as in Lemma 26. If y, z, K satisfy (20) then $y, z, \tilde{K}$ do. Additionally, if $t_i > 0$ then $z^tK_iz = \sigma_{\mathcal{K}^\oplus}(zz^t)$.*

Proof The first claim comes directly from $\tilde{K}z = Kz$. The second claim comes from $z^t\tilde{K}z = \sigma_{\mathcal{K}^\oplus}(zz^t)$ and a standard argument. ■

Since K does not appear in the loss optimality conditions (which depend only on z and y), we see that we can construct an optimal y, K, and z for the entire kernel learning problem where K is a convex combination of at most $n+1$ “atoms” of $\mathcal{K}$ that solve the problem. Our result is both substantially simpler than the result of Argyriou et al. (2005), and more general in that it applies to arbitrary convex loss functions: Argyriou et al. (2005)’s argument was essentially a saddle point argument which required differentiability of the loss.

Our work generalizes previous results in this field, in that we have shown that one may use an arbitrary kernel penalizer $F(K)$; previous work has used only δ functions. Exploring alternate penalizers and developing algorithms is a topic for future work.

9. Regularizations and Bias

In this section, we consider relaxing primal regularizations by means of infimal convolutions. The primary application is unregularized *bias* terms for kernel machines. In Section 9.1, we obtain very general results for biased regularizations, independent of particular loss functions and even particular regularizers. In Section 9.2, we use similar techniques to analyze leave-one-out computations for RKHS learning.

9.1 Learning with Bias

So far, we have considered regularization in an RKHS, deriving formulations of regularized least squares and an unbiased version of the support vector machine. In practice, the SVM is generally used with an unregularized *bias* term b (Poggio et al., 2001). This gives the regularization the property that $R(y + b1_n) = R(y)$—constant shifts of the regression values are free. This can be achieved by taking a base regularization such as $R = \lambda Q_{K^{-1}}$ and replacing it with $R' = \lambda Q_{K^{-1}} \star \delta_{1_n\mathbb{R}}$ where $1_n \in \mathbb{R}^n$ is the vector of all 1s. In this case, we will see that the subgradient relation for the regularization becomes

$$z \in \partial R'(y) \Leftrightarrow \begin{pmatrix} \lambda^{-1}K & 1_n \\ 1_n^t & 0 \end{pmatrix} \begin{pmatrix} z \\ b \end{pmatrix} = \begin{pmatrix} y \\ 0 \end{pmatrix}, \tag{21}$$

which relates y to both z and the bias parameter b.

We examine biased regularizations more generally. Without loss of generality, we cast this theorem in terms of an infimal convolution in the primal regularization, though by biconjugation, this lemma applies symmetrically to the primal and dual.

Lemma 28 (bias) *Let $C \subset \mathbb{R}^n$ be a closed convex set containing* 0. *Let $R : \mathbb{R}^n \to (-\infty, \infty]$ be ccp where $R' = R \star \delta_C$ is exact. Then $z \in \partial R'(y)$ iff $\exists c \in C$ such that*

$$\begin{aligned} \forall c' \in C, z^t(c'-c) &\leq 0 \\ R(y-c) - (y-c)^t z + R^*(z) &= 0. \end{aligned} \tag{22}$$

Proof By Theorem 15, $z \in \partial R'(y)$ iff $\exists c$ such that $z \in \partial \delta_C(c)$ and $z \in \partial R(y-c)$. The first condition is given by Equation 13, and the second condition is the generic condition of Theorem 6. ■

The first condition in Lemma 28 is a normality condition on z taking different forms depending on the geometry of C. It states that the hyperplane given by $H_{z,\sigma_C(z)}(y) = 0$ is a supporting hyperplane of C. For example, if the boundary of C were smooth, then this condition reduces to z being a normal to C at c. If $C = \mathbb{B}$, then $z = \lambda c$ for some $\lambda \geq 0$.

If C is a vector subspace, then $\sigma_C = \delta_{C^\perp}$, and (22) becomes the condition $z \in C^\perp$ with c an arbitrary element of C. Thus, it seems worthwhile to specialize to the case of vector subspaces.

Corollary 29 *Let $V_1 \subset V_2 \subset \mathbb{R}^n$ be vector subspaces. Let $R : \mathbb{R}^n \to (-\infty, \infty]$ be ccp. Let $R' = R \star \delta_{V_1} + \delta_{V_2}$. Then $R'^* = R^* \star \delta_{V_2^\perp} + \delta_{V_1^\perp}$ and (assuming exactness of the $\star$'s) $z \in \partial R'(y)$ iff $y \in V_2, z \in V_1^\perp$ and $z - a \in \partial R(y-b)$ for some $a \in V_2^\perp, b \in V_1$.*

Proof Let $R_1 = R \star \delta_{V_1}$ and $R_2 = R_1 + \delta_{V_2}$. By Lemma 28,

$$\begin{aligned} z_1 \in \partial R_1(y_1) &\Leftrightarrow z_1 \in V_1^\perp \text{ and } \exists b \in V_1 \text{ s.t. } z_1 \in \partial R_1(y_1 - b) \\ z_2 \in \partial R_2(y_2) &\Leftrightarrow y_2 \in V_2 \text{ and } \exists a \in V_2^\perp \text{ s.t. } z_2 - a \in \partial R_1(y_2). \end{aligned}$$

Thus,

$$\begin{aligned} z \in \partial R'^*(y) \quad &\Leftrightarrow \quad y \in V_2 \text{ and } \exists a \in V_2^\perp \text{ s.t. } z - a \in V_1^\perp \\ &\qquad\quad \text{and } \exists b \in V_1 \text{ s.t. } z - a \in \partial R(y-b) \\ &\Leftrightarrow \quad y \in V_2 \text{ and } z \in V_1^\perp \text{ and } \exists a \in V_2^\perp, b \in V_1 \text{ s.t. } z - a \in \partial R(y-b), \end{aligned}$$

where the last line is due to $V_2^\perp \subset V_1^\perp$. ■

Suppose we start with a regularization R, and we wish to allow a free constant bias in the y values. Our primal problem is:

$$\inf_{y,b} \left\{ R(y - b1_n) + \sum_{i=1}^n (1 - y_i Y_i)_+ \right\} = \inf_y \left\{ (R \star \delta_{1_n\mathbb{R}})(y) + \sum_{i=1}^n (1 - y_i Y_i)_+ \right\}.$$

In the dual, the regularization term becomes $(R^* + \delta_{(1_n\mathbb{R})^\perp})(z)$. We see that the dual regularizer $z \in (1_n\mathbb{R})^\perp$ can also be written as the constraint $\sum z_i = 0$. We have shown that for *any* regularization and loss function, allowing a free unregularized constant b in the y values induces a constraint $\sum_i z_i = 0$ in the dual problem. *Nothing else changes:* to obtain the standard "biased" SVM instead of the unbiased SVM we derived in Section 5.2.2, we change the primal regularizer from $Q_{K^{-1}}$ to $Q_{K^{-1}} \star \delta_{1_n\mathbb{R}}$, we change the dual regularizer from Q_K to $Q_K + \delta_{(1_n\mathbb{R})^\perp}$, or, equivalently, we add the constraint $\sum_i z_i = 0$ to the dual optimization problem. There is no need to take the entire dual again—the loss function is unchanged, and in the Fenchel formulation, the regularization and loss make separate contributions.

In terms of Corollary 29, the standard constant bias is obtained by choosing $V_1 = 1_n\mathbb{R}, V_2 = \mathbb{R}^n$, and therefore $a = 0$ and $R \star \delta_{V_1} + \delta_{V_2} = R \star \delta_{V_1}$. Assuming we are using an RKHS primal regularizer $Q_{K^{-1}}$, then $z \in \partial(Q_{K^{-1}} \star \delta_{V_1})(y-b)$ iff $b \in 1_n\mathbb{R}$ and $z \in \partial Q_{K^{-1}}(y-b)$, or, equivalently, if (21) holds.

We can incorporate the same bias into regularized least squares. Recalling that the loss optimality condition for RLS is $y+z=Y$, we can eliminate either y or z from the optimality condition, obtaining either a pure primal or pure dual formulation for the so-called *least squares SVM* (Suykens et al., 2002).

$$\begin{aligned}\begin{pmatrix}\lambda^{-1}K & 1\\ 1^t & 0\end{pmatrix}\begin{pmatrix}Y_i\\ b\end{pmatrix} &= \begin{pmatrix}I+\lambda^{-1}K & 1\\ 1^t & 0\end{pmatrix}\begin{pmatrix}y\\ 0\end{pmatrix}\\ \begin{pmatrix}I+\lambda^{-1}K & 1\\ 1^t & 0\end{pmatrix}\begin{pmatrix}z\\ b\end{pmatrix} &= \begin{pmatrix}Y_i\\ 0\end{pmatrix},\end{aligned}$$

Note that introducing biases via infimal convolutions with RKHS regularization preserves the representer property:

Corollary 30 *Let* $K = \begin{pmatrix}K_{NN} & K_{NM}\\ K_{MN} & K_{MM}\end{pmatrix} \in \mathbb{R}^{(n+m)\times(n+m)}$ *be symmetric positive definite where* $K_{NN} \in \mathbb{R}^{n\times n}$. *Let* $C \in \mathbb{R}^{n+m}$ *be a closed convex set containing 0, and let* C_N *be the projection of* C *onto the first N dimensions. For all* $y_1 \in \mathbb{R}^n$,

$$\inf_{y_2\in\mathbb{R}^m}(Q_{K^{-1}} \star \delta_C)(y_1,y_2) = (Q_{K_{NN}^{-1}} \star \delta_{C_N})(y_1),$$

where the minimizer is of the form $y_2 = K_{MN}K_{NN}^{-1}(y_1 - c_1) + c_2$ *where* $(c_1,c_2) \in C$.

Proof

$$\begin{aligned}\inf_{y_2\in\mathbb{R}^m}(Q_{K^{-1}} \star \delta_C)(y_1,y_2) &= \inf_{y_2\in\mathbb{R}^m}\inf_{(c_1,c_2)\in C} Q_{K^{-1}}((y_1,y_2)-(c_1,c_2))\\ &= \inf_{(c_1,c_2)\in C}\inf_{y_2\in\mathbb{R}^m} Q_{K^{-1}}((y_1,y_2)-(c_1,c_2))\\ &= \inf_{c_1\in C_N}\inf_{y_2'\in\mathbb{R}^m} Q_{K^{-1}}(y_1-c_1,y_2')\\ &= \inf_{c_1\in C_N} Q_{K_{NN}^{-1}}(y_1-c_1),\end{aligned}$$

where the last equality is an application of Corollary 24 (the statement about y_2 then follows since that $Q_{K^{-1}} \star \delta_C$ and $Q_{K_{NN}^{-1}} \star \delta_{C_N}$ are exact). ■

The minimizing y_2 will be unique if a *transversality* condition holds: $\forall (c_1,c_2),(c_1',c_2') \in C, c_1 = c_1' \Rightarrow c_1 = c_2'$. For the standard constant-term bias, this is clear and the recovery of y_2 reduces to $y_2 = K_{MN}K_{NN}^{-1}(y_1 - 1_n b) + 1_m b$.

9.2 Leave-one-out Computations

For small data sets, it is frequently of interest to do model selection via leave-one-out cross validation. Done naively, this involves solving n optimization problems, one for each size $n-1$ subset of the data, and testing $n-1$ functions at the remaining points. In this section, we relate the solution of the full problem (using all the data points) to the leave-one-out value.

We consider an $n+1$ point data set, and define $y = (y_0, y_N) \in \mathbb{R}^{n+1}$ in a regression problem with losses $v_i(y)$ for $0 \leq i \leq n$. The "full" optimization problem, with an RKHS regularization, is

$$\inf_{y \in \mathbb{R}^{n+1}} \left\{ \lambda Q_{K^{-1}}(y) + \sum_{i=0}^{n} v_i(y_i) \right\}.$$

We now show how we can compute a "leave one out" score as the auxiliary variable of a biased regularization of the form δ_C where $C = \{(y_0, 0) \in \mathbb{R}^{n+1}\}$. Consider the optimization,

$$\inf_{y \in \mathbb{R}^{n+1}} \left\{ \lambda (Q_{K^{-1}} \star \delta_C)(y) + \sum_{i=0}^{n} v_i(y_i) \right\} = \inf_{y \in \mathbb{R}^{n+1}, \beta \in \mathbb{R}} \left\{ \lambda Q_{K^{-1}}(y - \beta e_0) + \sum_{i=0}^{n} v_i(y_i) \right\}.$$

By Corollary 29, the right hand problem has the regularization optimality condition

$$\begin{pmatrix} \lambda^{-1} K & e_0 \\ e_0^t & 0 \end{pmatrix} \begin{pmatrix} z \\ \beta \end{pmatrix} = \begin{pmatrix} y \\ 0 \end{pmatrix},$$

Clearly, at optimality, $z_0 = 0$ and y_0 will assume a value with minimal loss: $v_0(y_0) = \inf_y v(y) \equiv \bar{v}$. With $z_0 = 0$, it is clear that the values $z_1, \ldots, z_n$ and $y_1, \ldots, y_n$ are an optimal solution to the restricted problem obtained by simply discarding (X_0, Y_0), and that therefore, $(\lambda^{-1} K z)_0$ is the value we would obtain if we solved the restricted problem and computed the "test value" at X_0. β is the leave-one-out "error" $y_0 - (\lambda^{-1} K z)_0$.

In the case of a square loss, the loss optimality conditions are $z + y = Y$, allowing us to eliminate y,

$$\begin{pmatrix} I + \lambda^{-1} K & e_0 \\ e_0^t & 0 \end{pmatrix} \begin{pmatrix} z \\ \beta \end{pmatrix} = \begin{pmatrix} Y \\ 0 \end{pmatrix},$$

and solve for β, Defining $G \equiv I + \lambda^{-1} K$, we have

$$\beta = \frac{e_0^t G^{-1} Y}{e_0^t G^{-1} e_0},$$

the standard formula for RLS leave-one-out errors.

10. Low-rank Approximation of Kernel Matrices

Given n data points, the kernel matrix K is n by n. For large n, it is impractical to compute with (or store) K. Consequently, there has been an interest in approximations to K. In this section, we consider a class of low rank approximations to K obtained by considering a vector subspace $V_0 \subset \mathbb{R}^n$, and defining $\tilde{K}$ as the lowest rank symmetric matrix such that $\tilde{K} V_0 = K V_0$. The primary example is the well-known Nyström approximation (Baker, 1977). When the Nyström kernel matrix approximation was first used in machine learning applications (Willams and Seeger, 2000), the suggested approach was to simply replace K with $\tilde{K}$; we refer to this approach as the *Nyström method*. We will show that the "natural" algorithm suggested by the Nyström *approximation* is in fact the *subset of regressors* method (Poggio and Girosi, 1990; Luo and Wahba, 1997), as opposed to the Nyström method.[9] In contrast, the Nyström method essentially makes an unwarranted (and incorrect) assumption that the representer theorem, which directly connects the dual variables z and the function expansion coefficients c in the standard RKHS case, still holds when we replace K with $\tilde{K}$. Our observation is consonant with empirical reports (Rifkin, 2002; Rasmusen and Williams, 2006) that the subset of regressors algorithm tends to outperform the Nyström algorithm. We note in passing that whereas previous algorithmic work with the Nyström approximation has mostly focussed on Gaussian Process regression or regularized least-squares, our results are all independent of the particular loss function chosen.

We first show that $Q_{\tilde{K}}$ has a simple characterization.

Lemma 31 *Let $K, \tilde{K} \in \mathbb{R}^{n\times n}$ with K symmetric positive definite and $\tilde{K}V_0 = KV_0$ with $\mathrm{rank}(\tilde{K}) = \dim V_0$. Then $Q_{\tilde{K}} = Q_K \star \delta_{(KV_0)^\perp}$.*

Proof For all $v \in V_0$ and $n \in Null(\tilde{K})$, $0 = n^t\tilde{K}v = n^tKv$, thus, since $\mathrm{rank}(\tilde{K}) = \dim V_0$, $Null(\tilde{K}) = (KV_0)^\perp$ and $V_0 \cap (KV_0)^\perp = \{0\}$.

Given $v \in V_0, n \in (KV_0)^\perp$,

$$\begin{aligned} Q_K(v+n) &= \frac{1}{2}(v+n)^tK(v+n) \\ &= Q_K(v) + Q_K(n) \\ Q_{\tilde{K}}(v+n) &= (v+n)^t\tilde{K}(v+n) \\ &= \frac{1}{2}v^t\tilde{K}v = \frac{1}{2}v^tKv = Q_K(v). \end{aligned}$$

In comparison

$$\begin{aligned} (Q_K \star \delta_{(KV_0)^\perp})(v+n) &= \inf_w \left\{ Q_K(v+n-w) + \delta_{(KV_0)^\perp}(w) \right\} \\ &= Q_K(v) + \inf_{w\in(KV_0)^\perp} \{Q_K(n-w)\} \\ &= Q_K(v). \end{aligned}$$

■

Lemma 31 implies

$$\begin{aligned} \inf_z \{\lambda^{-1}Q_{\tilde{K}}(z) + V^*(-z)\} &= \inf_z \left\{ \lambda^{-1}(Q_K \star \delta_{(KV_0)^\perp})(z) + V^*(-z) \right\} \\ &= \inf_{z\in\mathbb{R}^n, z'\in(KV_0)^\perp} \{\lambda^{-1}Q_K(z-z') + V^*(-z)\} \\ &= \inf_{z''\in\mathbb{R}^n} \left\{ \lambda^{-1}Q_K(z'') + \inf_{z'\in(KV_0)^\perp} V^*(-z''-z') \right\}. \qquad (23) \end{aligned}$$

We see that giving the regularization a non-trivial nullspace ($(KV_0)^\perp$) is equivalent to allowing an unregularized "bias" from that same space. Hence, we expect the resulting modified dual optimization problem to have a lower value than the original dual.

[9] To avoid confusion, we reiterate that the *Nyström approximation* is a low-rank approximation to K, denoted by $\tilde{K}$, while the *Nyström method* is obtained by simply replacing K with $\tilde{K}$ in an algorithm such as regularized least squares.

The modified primal problem, the Fenchel dual of (23), is

$$\inf_{y}\{\lambda Q_{K^{-1}}(y)+\delta_{KV_0}(y)+V(y)\} \quad = \quad \inf_{y\in KV_0}\{\lambda Q_{K^{-1}}(y)+V(y)\} \tag{24}$$

which is identical to the original optimization problem, but with a restricted domain, and hence a higher value, in general. This is to be expected, as any pair of Fenchel duals have inversely related infimal values. For the remainder of this section, the variables y and z refer, respectively to optimal solutions to the modified primal (24) and dual (23) problems.

So far, we have considered an arbitrary subspace V_0. We now specialize to the Nyström approximation, obtained by partitioning the data into two blocks N and M with $K=\begin{pmatrix}K_{NN} & K_{NM}\\ K_{MN} & K_{MM}\end{pmatrix}$, and taking $V_0=\begin{pmatrix}I\\0\end{pmatrix}\mathbb{R}^n$, in which case

$$\tilde{K}=\begin{pmatrix}K_{NN} & K_{NM}\\ K_{MN} & K_{MN}K_{NN}^{-1}K_{NM}\end{pmatrix}=\begin{pmatrix}I & 0\\ K_{MN}K_{NN}^{-1} & 0\end{pmatrix}\begin{pmatrix}K_{NN} & K_{NM}\\ K_{MN} & K_{MM}\end{pmatrix}.$$

It is obvious that $\tilde{K}V_0=KV_0$ and the expression on the right demonstrates that $\tilde{K}$ is rank n.

By Corollary 29, with $V_2=K\begin{pmatrix}I\\0\end{pmatrix}\mathbb{R}^n=\left(K^{-1}\begin{pmatrix}0\\I\end{pmatrix}\mathbb{R}^m\right)^{\perp}$ and $V_1=\{0\}$, the complementarity relation between y and z due to the regularization terms of the modified problems (24) and (23) is

$$\begin{pmatrix}z\\0\end{pmatrix}=\lambda\begin{pmatrix}K^{-1} & K^{-1}\begin{pmatrix}0\\I\end{pmatrix}\\ \begin{pmatrix}0 & I\end{pmatrix}K^{-1} & 0\end{pmatrix}\begin{pmatrix}y\\ \alpha\end{pmatrix}$$

or, equivalently, $\exists r\in\mathbb{R}^n$ such that

$$y \quad = \quad \lambda^{-1}K\begin{pmatrix}r\\0\end{pmatrix} \tag{25}$$

and

$$\lambda^{-1}Kz \quad = \quad y+\begin{pmatrix}0\\ \alpha\end{pmatrix} \tag{26}$$

$$\lambda^{-1}\tilde{K}z \quad = \quad y, \tag{27}$$

where (27) is obtained from (26) by multiplying the first equation on the left by $\begin{pmatrix}I & 0\\ K_{MN}K_{NN}^{-1} & 0\end{pmatrix}$ and observing that $\tilde{K}\begin{pmatrix}r\\0\end{pmatrix}=K\begin{pmatrix}r\\0\end{pmatrix}$ by construction.

In standard Tikhonov regularization in an RKHS, the complementarity equation from the regularization, $y=\lambda^{-1}Kz$, and the representer theorem tells us that we can obtain an optimizing function using $c=\lambda^{-1}z$. Once we have replaced K with $\tilde{K}$, this connection no longer holds, and we must decide how to generalize a minimizer of (23) or (24) into a function that can be used to predict the values at future points. The Nyström method, suggested in Willams and Seeger (2000) is to use the function $f_d(x)\equiv\sum_{i=1}^{n+m}\lambda^{-1}z_ik(x,X_i)$, essentially "pretending" that the connection between c

and z is still valid. An alternate choice is to use r, yielding $f_p(x) \equiv \sum_{i=1}^n \lambda^{-1} r_i k(x, X_i)$.[10] By (26), $f_p(X_i) = f_d(X_i)$ when $1 \le i \le n$, while $f_d(X_i) = f_p(X_i) + \alpha_{i-n}$ when $n < i \le n+m$—the two regressors are related, taking the same values at the first n points but generally different values on the last m points.

By eliminating y for r in the modified primal problem (24), we see that f_p is the solution of the *subset of regressors* method, where we construct a function using only those coefficients associated with points in N:

$$\inf_{y \in KV_0} \{\lambda Q_{K^{-1}}(y) + V(y)\} \quad = \quad \inf_{r \in \mathbb{R}^n} \left\{ \lambda^{-1} Q_{K_{NN}}(r) + V\left(\lambda^{-1} \begin{pmatrix} K_{NN} \\ K_{MN} \end{pmatrix} r \right) \right\}.$$

In this sense, we can see that the subset of regressors method is associated with the modified problems (24) and (23). Additionally, the function f_p will recover the y values optimizing (24) at all $n+m$ training points. In contrast, if we use the Nyström method on the last m training points, we will *not* recover the last m values—they will differ by α.

We have shown that the subset of regressors method corresponds in a natural way to the modified primal and dual problems obtained by replacing K with the Nyström approximation $\tilde{K}$, whereas the Nyström method makes the additional unwarranted assumption that it is still a good idea to construct a function to classify future points using $c = \lambda^{-1} z$. We can also derive an interesting relationship between the functions f_p and f_d.

We first note that by (25), y and $\begin{pmatrix} 0 \\ \alpha \end{pmatrix}$ are "K^{-1} orthogonal": $y^t K^{-1} \begin{pmatrix} 0 \\ \alpha \end{pmatrix} = \begin{pmatrix} r \\ 0 \end{pmatrix}^t \begin{pmatrix} 0 \\ \alpha \end{pmatrix} = 0$. As a consequence, the quadratic form $Q_{K^{-1}}$ "distributes" over y and $\begin{pmatrix} 0 \\ \alpha \end{pmatrix}$:

$$\begin{aligned} Q_{K^{-1}}(\lambda^{-1} K z) &= Q_{K^{-1}}\left(y + \begin{pmatrix} 0 \\ \alpha \end{pmatrix} \right) \\ &= Q_{K^{-1}}(y) + y^t K^{-1} \begin{pmatrix} 0 \\ \alpha \end{pmatrix} + Q_{K^{-1}} \begin{pmatrix} 0 \\ \alpha \end{pmatrix} \\ &= Q_{K^{-1}}(y) + Q_{K^{-1}} \begin{pmatrix} 0 \\ \alpha \end{pmatrix}. \end{aligned}$$

Furthermore, by (26),

$$\begin{aligned} \begin{pmatrix} 0 \\ \alpha \end{pmatrix}^t z &= \begin{pmatrix} 0 \\ \alpha \end{pmatrix}^t \left(\begin{pmatrix} r \\ 0 \end{pmatrix} + \lambda K^{-1} \begin{pmatrix} 0 \\ \alpha \end{pmatrix} \right) \\ &= \lambda \begin{pmatrix} 0 \\ \alpha \end{pmatrix}^t K^{-1} \begin{pmatrix} 0 \\ \alpha \end{pmatrix} \\ &= 2\lambda Q_{K^{-1}} \begin{pmatrix} 0 \\ \alpha \end{pmatrix}. \end{aligned}$$

We now consider comparing the functions f_d and f_p. Recall that the values of f_d at all training points are equal to $\lambda^{-1} K z$ while those of f_p are y. Because y and z are optimal solutions to the modified

[10] We chose the name f_d and f_p because the two functions seem to us "suggested" by the modified dual and primal problems, respectively.

primal and dual problems, $-z \in \partial V(y)$. Since $-z$ is a subgradient of V at y, $V(y) - z^t(y'-y) \leq V(y')$ for all $y' \in \mathbb{R}^n$. Setting $y' = \lambda^{-1}Kz$, and noting that $\lambda^{-1}Kz - y = \begin{pmatrix} 0 \\ \alpha \end{pmatrix}$, by (26), we have,

$$\begin{aligned} V(y) - V(y') \leq z^t(y'-y) &= z^t \begin{pmatrix} 0 \\ \alpha \end{pmatrix} = 2\lambda \left(Q_{K^{-1}}(\lambda^{-1}Kz - y) \right) \\ &= 2\lambda \left(Q_{K^{-1}}(y') - Q_{K^{-1}}(y) \right) = 2\lambda Q_{K^{-1}} \begin{pmatrix} 0 \\ \alpha \end{pmatrix}. \end{aligned}$$

Recall that y was obtained by minimizing $V(y) + \lambda Q_{K^{-1}}(y)$ over KV_0. Comparing the function values y and y' of the regressors f_p and f_d respectively, we see that f_d pays a higher regularization cost as compared to f_p, $Q_{K^{-1}}(y') - Q_{K^{-1}}(y) = Q_{K^{-1}} \begin{pmatrix} 0 \\ \alpha \end{pmatrix}$. Furthermore, the difference in loss at the last m points in favor of f_d is bounded by 2λ times this difference—it is not possible for f_d to have arbitrarily smaller loss. In practice, we often find that the Nyström method produces *worse* function values at the m points than the subset of regressors does. This provides further evidence that the subset of regressors method should be preferred to the Nyström method.

11. Discussion and Future Work

We have introduced a new framework for thinking about optimization in learning theory by combining two key elements: Fenchel duality and value regularization. Fenchel duality allows us to analyze the optimality conditions for regularization and loss terms separately, and combine these optimality conditions to get complete optimality conditions for machine learning schemes. Value regularization makes it easy to reason about optimization problems over training and testing points simultaneously, and isolates the RKHS kernel to the regularization term. We have used the framework to gain new insights into several topics in machine learning, including the representer theorem, learning a kernel matrix, biased regularizations, and low-rank approximations to kernel matrices. There remain several interesting open questions.

In Sections 7 and 8, we showed that both supervised learning in an RKHS and learning the kernel matrix had a representer theorem that caused the inductive and natural transductive algorithms to make the same prediction. It would be interesting to explore this idea further, characterizing formally what sorts of conditions give rise to agreement between inductive and transductive algorithms.

The framework developed in this article is specialized to the case where the predicted outputs y are real-valued scalars. It is frequently of interest to make predictions on objects with more structure: examples include multiclass classification, ranking, and classification of sequences. The extension of the value regularization framework to these problems is in principle straightforward, but many details remain to be worked out.

Finally, although this article has focussed on the analysis of existing schemes, one may ask whether a value regularization perspective can lead to the development of new learning algorithms. It seems unlikely that value regularization in an RKHS will lead directly to improved algorithms (for example, a better SVM solver), because the inverse kernel matrix K^{-1} appearing in the primal is a computationally inconvenient object. However, we might imagine designing regularizers for which value-based optimization is computationally effective, such as a transductive algorithm with a regularizer that directly imposes some non-RKHS smoothness over the input space. We are actively working to design and implement such algorithms.

Acknowledgments

We thank Ben Recht, Tomaso Poggio, Rob Freund, Adrian Lewis and Roger Wets for helpful discussions and suggestions. We thank Gustavo Santos and Tim Sears for finding many technical errors in earlier versions of this work. We thank our anonymous reviewers for helpful suggestions.

References

Yasemin Altun and Alex Smola. Unifying divergence minimization and statistical inference via convex duality. In *Proceedings of the 19th Annual Conference on Lwearning Theory*, 2006.

Yasemin Altun, David McAllester, and Mikhail Belkin. Maximum margin semi-supervised learning for structured variables. In *Neural Information Processing Systems*, 2005.

Andreas Argyriou, Charles A. Micchelli, and Massimiliano Pontil. Learning convex combinations of continuously parametrized basic kernels. In *Proceedings of the 18th Annual Conference on Learning Theory*, 2005.

Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

Christopher T. H. Baker. *The Numerical Treatment of Integral Equations*. Clarendon press, 1977.

Mokhtar S. Bazaraa, Hanif D. Sherali, and C. M. Shetty. *Nonlinear Programming: Theory and Algorithms*. Wiley-Interscience, 2nd edition, 1993.

Mikhail Belkin and Partha Niyogi. Semi-supervised learning on riemannian manifolds. *Machine Learning*, 56:209–239, 2003.

Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from examples. Technical Report TR-2004-06, University of Chicago, 2004.

Yoshua Bengio, Jean-Francois Paiement, Pascal Vincent, Olivier Delalleau, Nicolas Le Roux, and Marie Ouimet. Out-of-sample extensions for LLE, Isomap, MDS, eigenmaps and spectral clustering. In *Neural Information Processing Systems*, 2003.

Jonathan M. Borwein and Adrian S. Lewis. *Convex Analysis and Nonlinear Optimization*. CMS Books in Mathematics. Springer, 2000.

Scott Shaobing Chen, David L. Donoho, and Michael A. Saunders. Atomic decomposition by basis pursuit. Technical Report 479, Stanford University Department of Statistics, 1995.

Corinna Cortes and Vladimir Vapnik. Support vector networks. *Machine Learning*, 20:1–25, 1995.

Nello Cristianini and John Shaw-Taylor. *An Introduction To Support Vector Machines*. Cambridge University Press, 2000.

Miroslav Dudík and Robert E. Schapire. Maximum entropy distribution estimation with generalized regularization. In *Proceedings of the 19th Annual Conference on Lwearning Theory*, pages 123–138, 2006.

Theodoros Evgeniou, Massimiliano Pontil, and Tomaso Poggio. Regularization networks and support vector machines. *Advances In Computational Mathematics*, 13(1):1–50, 2000.

Federico Girosi. An equivalence between sparse approximation and support vectom̊achines. *Neural Computation*, 10:1455–1480, 1998.

Tommi S. Jaakkola and David Haussler. Probabilistic kernel regression models. In *Proceedings of the 1999 Conference on AI and Statistics*. Morgan Kaufmann, 1999.

Gert R. G. Lanckriet, Nello Cristianini, Peter Bartlett, Laurent El Ghaoui, and Michael I. Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, 5:24–72, 2004.

Zhen Luo and Grace Wahba. Hybrid adaptive splines. *Journal of the American Statistical Society*, 92:107–116, 1997.

Charles A. Micchelli. Interpolation of scattered data: distance matrices and conditionally positive functions. *Constructive Approximation*, 2(1):11–22, 1986.

Tomaso Poggio and Federico Girosi. Networks for approximation and learning. *Proceedings of the IEEE*, 78(9):1481–1497, September 1990.

Tomaso Poggio, Sayan Mukherjee, Ryan M. Rifkin, Alex Rakhlin, and Alessandro Verri. b. In *Proceedings of the Conference on Uncertainty in Geometric Computations*, 2001.

Ali Rahimi, Benjamin Recht, and Trevor Darrell. Learning appearance manifolds from video. In *Computer Vision and Pattern Recognition*, 2005.

Carl Edward Rasmusen and Chris Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

Ben Recht. Unpublished phd thesis. 2006.

Ryan M. Rifkin. *Everything Old Is New Again: A Fresh Look at Historical Approaches to Machine Learning*. PhD thesis, Massachusetts Institute of Technology, 2002.

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.

Vikas Sindhwani, Partha Niyogi, and Mikhail Belkin. Beyond the point cloud: from transductive to semi-supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, 2005.

Johan A. K. Suykens, Tony Van Gestel, Jos De Brabanter, Bart De Moor, and Joos Vandewalle. *Least Squares Support Vector Machines*. World Scientific, 2002.

Andrei N. Tikhonov and Vasilii Y. Arsenin. *Solutions of Ill-posed problems*. W. H. Winston, Washington D.C., 1977.

Vladimir Vapnik. *Statistical Learning Theory*. John Wiley & Sons, 1998.

Grace Wahba. *Spline Models for Observational Data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial & Applied Mathematics, 1990.

Christopher K. I. Willams and Matthias Seeger. Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems*, 2000.

Ji Zhu, Saharon Rosset, Trevor Hastie, and Rob Tibshirani. 1-norm support vector machines. In *Neural Information Processing Systems*, 2003.