Generalized Kernel Classification and Regression

Jason Palmer and Kenneth Kreutz-Delgado

Department of Electrical and Computer Engineering
University of California San Diego, La Jolla, CA 92093, USA
japalmer@ucsd.edu,kreutz@ece.ucsd.edu

Abstract. We discuss kernel-based classification and regression in a general context, emphasizing the role of convex duality in the problem formulation. We give conditions for the existence of the dual problem, and derive general globally convergent classification and regression algorithms for solving the true (i.e. hard-margin or rigorous) dual problem without resorting to approximations.

1 Introduction

Kernel methods perform perform optimization in Hilbert space by means of a finite dimensional dual problem. The conditions for the formulation of the dual problem essentially determine what we can "do in feature space", i.e. which optimization problems can be solved involving vectors in Hilbert space. Thus convex analysis plays a major role in the theory of kernel methods. The primary purpose of this paper is to derive general algorithms for kernel-based classification and regression by considering the problem from the viewpoint of convex analysis as represented by [8]. We give a summary of the relevant background in the appendix.

In the literature, kernel machines for classification and regression are generally presented in the standard forms given in [11]. Alternatives are occasionally considered, but they are the generally derived on an individual basis rather than by exploiting the powerful and elegant framework of convex analysis. The general classification case is considered in [3], where generalized kernel machines are defined, and an algorithm is developed to solve the primal classification problem for the case of the logistic link function. In this paper we expand on these ideas and develop a general convex analysis framework for both classification and regression. We believe that consideration of more general formulations of kernel methods aids in the understanding of those methods that are commonly used, e.g. by yielding greater insight into the meaning of various parameters. Considering the general case also yields insight into the relationship between kernel methods that are formulated as an optimization on Hilbert space, and the methods that are formulated directly as an optimization of the dual variables, as in [10] and [2].

We derive here general algorithms for kernel classification and regression. In each case we find that a coordinate descent algorithm is possible using the Lagrangian. In the classification case, we discuss the general form of the dual problem, and mention a simple extension of the standard SVM that eliminates the need for the upper bound constraint. In the case of regression, we discuss the quantitative and qualitative characteristics of conjugate functions. We define the concept of square-concavity and derive inequalities used to guarantee descent for certain loss functions, and we explain the relationship of this concept to the curvature properties of primal and dual objective problems.

Our development is similar to that in [11], but we work in greater generality, which allows a more elegant and intuitive formulation of the problems. For simplicity in the exposition of convex duality ideas, we generally assume finite dimensional vectors, and write $\mathbf{x}^T\mathbf{y}$ rather than $\langle\mathbf{x},\mathbf{y}\rangle$ for the inner product, though we usually have Hilbert space in mind. The extension is generally immediate by replacing all instances of $\mathbf{x}^T\mathbf{y}$ in vectors or matrices by $\langle\mathbf{x},\mathbf{y}\rangle$.

2 Classification

Consider the two class problem, with data and class label pairs $(\mathbf{x}_1, y_1)$, ..., $(\mathbf{x}_N, y_N)$, $y_i \in \{-1, 1\}$. Let the decision hyperplane be defined by,

$$\mathbf{w}^T\mathbf{x} - b = 0\,, \qquad \mathbf{w}^T\mathbf{w} = 1$$

The signed distance of the point $\mathbf{x}_i$ to the hyperplane is $y_i\,(\mathbf{w}^T\mathbf{x}_i - b\,) \equiv z_i$.

2.1 Probabilistic model

In the probabilistic model, the random variables Y and X correspond to class label and data point, and $Z = Y(\mathbf{w}^T X - b)$. The posterior class likelihood given a data point $\mathbf{x}$ and the parameters $\mathbf{w}$ and b, is modelled as a function of Z only, that is $P(Y\!=y|X\!=\mathbf{x}) = P(Y\!=y|Z\!=z)$. In this model, the probability of a particular class labelling of N independent samples is,

$$\prod_{i=1}^{N} P\big(y_i|\mathbf{x}_i, \mathbf{w}, b\big) = \prod_{i=1}^{N} F\big(y_i\,(\mathbf{w}^T\mathbf{x}_i - b\,)\big) = \prod_{i=1}^{N} F(z_i) \tag{1}$$

where $F(z) \equiv P(Y\!=y|Z\!=z)$. Let $\mathbf{y} = [y_1 \cdots y_N]^T$ be the vector of class labels. We attempt to find the parameters $\mathbf{w}$ and b that maximize the log probability of the samples,

$$l(\mathbf{w}, b\,|\,\mathbf{y}, \{\mathbf{x}_i\}) \;=\; \sum_{i=1}^{N} \log F(z_i) \;\equiv\; \sum_{i=1}^{N} f(z_i)$$

subject to the constraints $\mathbf{w}^T\mathbf{w} = 1$ and $y_i\,(\mathbf{w}^T\mathbf{x}_i - b\,) = z_i$. Now let $\mathbf{\Phi}$ be the matrix with $y_i\mathbf{x}_i$ in the ith column, and let $\mathbf{z} = [z_1 \cdots z_N]^T$ be the vector of signed sample distances. Then the constraint on $\mathbf{z}$ can be written $\mathbf{z} = \mathbf{\Phi}^T\mathbf{w} - b\,\mathbf{y}$. Defining $f(\mathbf{z}) \equiv \sum_i f(z_i)$, the problem becomes,

$$\max_{\mathbf{z},\mathbf{w},b} f(\mathbf{z}) \quad s.t. \quad \mathbf{z} = \mathbf{\Phi}^T\mathbf{w} - b\,\mathbf{y}\,, \quad \mathbf{w}^T\mathbf{w} \le 1 \tag{2}$$

If F (our model of the posterior class distribution) is log-concave, so that f is concave, then (2) is a concave program (see the Appendix). If $\sup_{\mathbf{z}} f(\mathbf{z}) < \infty$, then since the Slater condition is obviously satisfied, Theorem 2 applies. If we define the Lagrangian,

$$L(\mathbf{z}, \mathbf{w}, b, \lambda, \mu) \;=\; f(\mathbf{z}) \;+\; \lambda^T \big(\boldsymbol{\Phi}^T \mathbf{w} - b\, \mathbf{y} - \mathbf{z}\big) \;+\; \mu\,(1 - \mathbf{w}^T \mathbf{w})$$

then by Theorem 2, we have,

$$\begin{aligned} \max_{\mathbf{z},\mathbf{w},b} \min_{\lambda} \min_{\mu \geq 0} \;\; L(\mathbf{z}, \mathbf{w}, b, \lambda, \mu) \;\;&=\;\; \min_{\lambda} \min_{\mu \geq 0} \max_{\mathbf{z},\mathbf{w},b} \;\; L(\mathbf{z}, \mathbf{w}, b, \lambda, \mu) \\ = \min_{\lambda} \min_{\mu \geq 0} \max_{b} \;\; \mu \left[\max_{\mathbf{w}} \; \frac{1}{\mu} \lambda^T \boldsymbol{\Phi}^T \mathbf{w} - \mathbf{w}^T \mathbf{w} \right] &- \left[\min_{\mathbf{z}} \; \lambda^T \mathbf{z} - f(\mathbf{z}) \right] - b\, \lambda^T \mathbf{y} + \mu \end{aligned}$$

Identifying the conjugate functions and treating b as a Lagrange multiplier, the problem becomes,

$$\min_{\lambda \in\, \partial f} \; \min_{\mu \geq 0} \;\; \frac{1}{4\mu}\, \lambda^T \boldsymbol{\Phi}^T \boldsymbol{\Phi}\, \lambda - f^*(\lambda) + \mu \quad s.t. \quad \lambda^T \mathbf{y} = 0 \tag{3}$$

Solving for and substituting the optimal $\mu_{\min} = \frac{1}{2} \left(\lambda^T \boldsymbol{\Phi}^T \boldsymbol{\Phi}\, \lambda \right)^{1/2}$, we can write the problem as an optimization over λ only,

$$\min_{\lambda \in\, \partial f} \;\; \left(\lambda^T \boldsymbol{\Phi}^T \boldsymbol{\Phi}\, \lambda \right)^{1/2} - f^*(\lambda) \quad s.t. \quad \lambda^T \mathbf{y} = 0 \tag{4}$$

This is essentially the method used in [11, §10.2.1], though the following framework is employed there.

2.2 Generalized optimal hyperplane model

The decision hyperplane can also be determined according to the generalized optimal hyperplane framework [11, §10.2.1], in which we attempt to find the hyperplane that (i) makes the fewest errors, and (ii) maximizes the margin of the correctly classified samples. The first criterion does not lead to overfitting, as according to the Structural Risk Minimization principle [11, §6.1.1] generalization ability is controlled by a parameter h, which determines a structure of sets of functions of constant VC dimension, or resolution. For fixed h, we find the function that makes the fewest errors. Among all hyperplanes that attain the minimum number of errors, we choose the one that maximizes the margin.

We would like to determine a computationally feasible optimization problem in accordance with these requirements. We shall assume the form given by (2), and determine the function $f(z)$. The variable z is the distance of the sample from the decision hyperplane. If z is positive, the sample is on the "correct" side of the hyperplane, i.e. is classified correctly, and if z is negative, the sample is misclassified. The function $f(z)$ represents the log probability that the sample is correctly classified given its distance from the hyperplane, but it should be viewed here more as a negative cost function that we attempt to determine subject to our requirements.

Our first priority is to make as few errors as possible. This suggests that we make the cost of misclassification much greater than the cost of correct classification. Our second priority is to maximize the margin. To achieve this, our cost function should depend only on the points on or near the margin, i.e. those points among the correctly classified that are closest to the hyperplane. Small perturbations of the points that are incorrectly classified, or of the points that are "inside" the margin, should leave the optimal hyperplane unchanged. This suggests that we make the cost function flat over $z < 0$ and $z > \delta$, where $\delta > 0$ is the margin.

Unfortunately, we do not know the value of the optimal margin. If we set δ higher than the optimal margin, then correctly classified data will affect the hyperplane. If we set δ lower than the optimal margin, then the optimal hyperplane will not be unique, and we are likely to end up with a margin that is less than optimal for one of the classes. This issue can be addressed by optimizing δ as well. We add a term to the primal cost function that rewards increasing δ enough to define the optimal hyperplane for a given number of errors as that with the largest margin. If we make this additional term linear, $\nu\,\delta$, then we have the ν-SVM of [9]. As ν is increased from zero, the margin will increase to contain more samples.

Another problem with our ideal cost function is that it is not concave. Since we ultimately have kernel methods in mind, we are limited to functions f that are concave. Thus the closest we can come to flat over $z < 0$ is linear. The consequence is that the optimal hyperplane will depend on *how* misclassified the misclassified samples are. Linear decay with magnitude can be interpreted as minimizing this dependence given the concavity requirement, making the cost function "robust" to outliers in the sense of robust loss functions in regression. The benefits of asymptotic linearity of the log probability however come at a cost to optimization complexity. See Example 3 below.

Examples

1. The perturbed step function $F(x) = 0$, $x < \delta$, $F(x) = 1$, $x \geq \delta$, has $f(x) = -\infty$, $x < \delta$, $f(x) = 0$, $x \geq 0$, and

$$f^*(\lambda) = \begin{cases} \delta\lambda & \lambda \geq 0 \\ -\infty & \lambda < 0 \end{cases}$$

   The supremum of this function is ∞ so Theorem 2 does not apply.

2. Vapnik's hard margin generalization of the optimal hyperplane can be formulated,

$$\min_{0 \leq \lambda \leq C} \; \left(\lambda^T \mathbf{\Phi}^T \mathbf{\Phi}\,\lambda\right)^{1/2} - \delta \sum_i \lambda_i \quad s.t. \quad \lambda^T \mathbf{y} = 0 \tag{5}$$

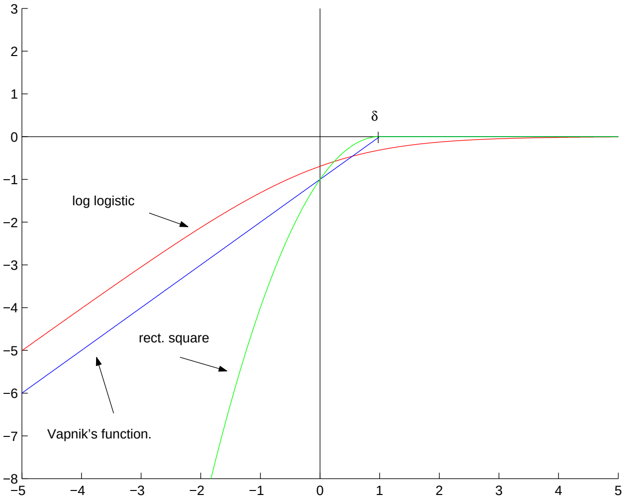

Fig. 1. Some log probability functions. The log logistic and Vapnik functions are asymptotically linear, and thus robust to misclassification, but as range ∂f is bounded above, so is domf^*, the domain of the dual problem

Identifying the conjugate function,

$$f^*(\lambda) = \begin{cases} \delta\lambda & 0 \leq \lambda \leq C \\ -\infty & \text{otherwise} \end{cases}$$

we can determine the corresponding functions $F(x) = \exp(-C(\delta - x)_+)$ and $f(x) = -C(\delta - x)_+$.

3. The rectified Gaussian, $F(x) = e^{-(x-\delta)^2/2\sigma^2}$ for $x \leq \delta$, $F(x) = 1$ for $x > \delta$, has $f(x) = -(x-\delta)_+^2/2\sigma^2$, and,

$$f^*(\lambda) = \begin{cases} -\frac{1}{2}\sigma^2\lambda^2 + \delta\,\lambda & \lambda \geq 0 \\ -\infty & \lambda < 0 \end{cases}$$

The fact that range $\partial f = $ domf^* is unbounded above simplifies the optimization problem (6). This function is less robust to mislabelled samples than the asymptotically linear functions. The optimization problem corresponding to this example differs from that in Example 2 only by adding a term $\sigma^2\mathbf{I}$ to the matrix $\mathbf{K}$. The benefit of this addition is that for any $\sigma > 0$, our optimization problem becomes,

$$\min_{\lambda\geq 0}\min_{\mu\geq 0} \quad \frac{1}{2}\lambda^T\left(\frac{1}{2\mu}\mathbf{K} + \sigma^2\mathbf{I}\right)\lambda - \delta\,\mathbf{e}^T\lambda + \mu \quad s.t. \quad \lambda^T\mathbf{y} = 0 \tag{6}$$

which makes the computational load substantially lighter by removing the upper bound constraint, and adding a regularizing ridge penalty term. This algorithm works well for small to medium sized problems, but as it generally yields non-sparse solutions, it is not suitable for large scale problems.

4. The logistic function, $F(x) = (1+\exp(-x))^{-1}$, has $f(x) = -\log(1+\exp(-x))$. The conjugate of the log logistic has the same form as the Shannon entropy,

$$f^*(\lambda) = -\lambda \log \lambda - (1-\lambda)\log(1-\lambda) \quad 0 \leq \lambda \leq 1$$

The logistic is asymptotically linear, like Vapnik's functions, making it robust. It also yields probabilistic information for all samples since it is strictly increasing on $(-\infty, \infty)$. However, given that the dual problem maximizes the entropy function, the solutions are decidedly non-sparse.

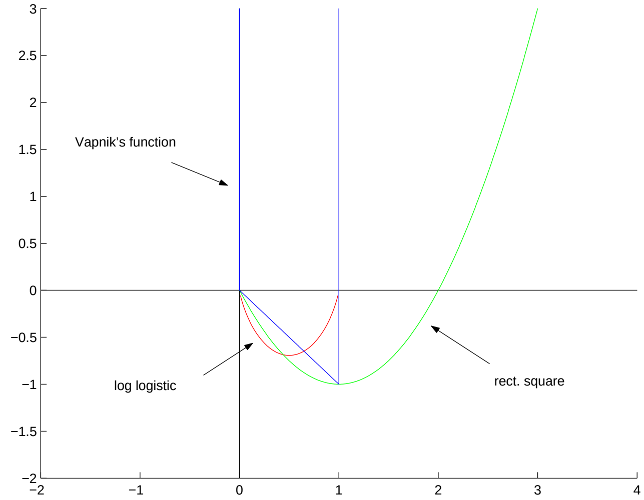

Fig. 2. The negative concave conjugates of the functions in Fig. 1.

The matrix $\mathbf{K} \equiv \mathbf{\Phi}^T\mathbf{\Phi}$ consists only of inner products of columns of $\mathbf{\Phi}$. This is exploited In kernel machines to solve primal problems in Hilbert space by means of a dual problem of finite dimension N. The dual formulation is possible only when the function F is log concave.

Global convergence by coordinate descent. The SVM is normally constructed using a simplified optimization problem, referred to as the soft-margin SVM, which is easier to solve computationally. We can however find the hard-margin SVM in a computationally straightforward way by performing coordinate descent on (6), i.e. by alternately minimizing (6) with respect to λ and μ. Given

$\mu = \mu_{k-1}$, the algorithm,

$$\lambda_k \;\leftarrow\; \arg\min_{\lambda \in \partial f} \; \frac{1}{4\mu_{k-1}} \, \lambda^T \mathbf{K} \, \lambda - f^*(\lambda) \quad s.t. \quad \lambda^T \mathbf{y} = 0 \tag{7}$$

$$\mu_k \;\leftarrow\; \arg\min_{\mu \geq 0} \; \frac{1}{4\mu} \, \lambda_k^T \mathbf{K} \, \lambda_k + \mu \;=\; \frac{1}{2} \, \|\lambda_k\|_{\mathbf{K}} \tag{8}$$

is a sequential quadratic programming (SQP) algorithm that monotonically decreases the objective function in (6).

An alternative formulation. Alternatively, we can rewrite (6) putting μ^2 for μ and defining $\alpha \equiv \lambda/\mu$ to get,

$$\min_{\alpha \in \partial f} \min_{\mu \geq 0} \; \frac{1}{4} \, \alpha^T \mathbf{K} \, \alpha - f^*(\mu\alpha) + \mu^2 \quad s.t. \quad \alpha^T \mathbf{y} = 0 \tag{9}$$

For Vapnik's SVM, we minimize $\frac{1}{4} \, \alpha^T \mathbf{K} \, \alpha - \delta\mu \mathbf{e}^T \alpha + \mu^2$ subject to $0 \leq \alpha \leq 1$, $\mu \geq 0$, and $\alpha^T \mathbf{y} = 0$. Optimizing over μ, we get,

$$\min_{0 \leq \alpha \leq 1} \; \alpha^T \big(\mathbf{K} - \delta^2 \mathbf{e}\mathbf{e}^T \big) \alpha \quad s.t. \quad \alpha^T \mathbf{y} = 0 \tag{10}$$

We emphasize that we have not made any approximations. Apparently (10) is equivalent to (5). In order to have $\alpha > 0$, δ must be large enough such that $\mathbf{K} - \delta \mathbf{e}\mathbf{e}^T$ is indefinite. Then the optimal α will lie in the direction of the subspace determined by the smallest eigenvalues, but constrained to be non-negative.. The sparsity in the solution of the SVM results from the fact that the optimal λ lies on an "edge" of the "cube", $0 \leq \alpha_i \leq \delta$, $i = 1, \ldots, N$.

3 Regression

Let $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$ be point and function value pairs observed from a nonlinear function $y(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}$. Consider a linear approximation of y given by,

$$\mathbf{w}^T \mathbf{x} + b = y \, , \qquad \mathbf{w}^T \mathbf{w} \leq A^2$$

Let $\mathbf{\Phi}$ be the matrix with $\mathbf{x}_i$ in column i, and let $z_i = \mathbf{w}^T \mathbf{x}_i + b - y_i$ denote the residuals. We define a probabilistic model in which z is random with symmetric density $p(z)$, and attempt to find parameters $\mathbf{w}$ and b that minimize the negative log likelihood $-\sum \log p(\mathbf{z}_i) \equiv d(\mathbf{z})$ of the samples,

$$\min_{\mathbf{z}, \mathbf{w}, b} \; d(\mathbf{z}) \quad s.t. \quad \mathbf{z} = \mathbf{\Phi}^T \mathbf{w} + b \, \mathbf{e} - \mathbf{y} \, , \quad \mathbf{w}^T \mathbf{w} \leq A^2$$

where $\mathbf{e}$ is the vector of all 1's. This is a concave program, similar to that obtained for the classification problem. For the Lagrangian, we have,

$$L(\mathbf{z}, \mathbf{w}, b, \lambda, \mu) \;=\; d(\mathbf{z}) \;+\; \lambda^T \big(\mathbf{y} - \mathbf{\Phi}^T \mathbf{w} - b \, \mathbf{e} - \mathbf{z} \big) \;+\; \mu \, (\mathbf{w}^T \mathbf{w} - A^2)$$

Applying Theorem 2 again, and rewriting the dual as a minimization, we get the equivalent problem,

$$\min_{\lambda \in \,\partial f} \min_{\mu \geq 0} \quad \frac{1}{4\mu}\, \lambda^T \mathbf{K}\, \lambda \,-\, \mathbf{y}^T \lambda \,+\, d^*(\lambda) \,+\, A^2 \mu \quad s.t. \quad \mathbf{e}^T \lambda = 0 \tag{11}$$

Substituting the optimal $\mu_{\min} = \frac{1}{2A} \|\lambda\|_{\mathbf{K}}$, we get,

$$\min_{\lambda \in \,\partial f} \; A \|\lambda\|_{\mathbf{K}} \,-\, \mathbf{y}^T \lambda \,+\, d^*(\lambda) \quad s.t. \quad \mathbf{e}^T \lambda = 0$$

The requirement for formulation of the dual problem in the case of regression, is that $d(z)$ be convex and symmetric. As in the classification case, we can solve (11) by coordinate descent in λ and μ for suitable $d(z)$. The regression case is somewhat simpler, however, given the assumed symmetry of $d(z)$, and becomes particularly simple when range $\partial f = \mathbb{R}$ so that there are no inequality constraints.

Examples

1. The indicator function $d(z) = 0,\, |z| \leq \delta,\; d(z) = \infty,\, |z| > \delta$, has,

$$d^*(\lambda) = \delta|\lambda| \quad \lambda \in \mathbb{R}$$

2. Vapnik's ϵ-insensitive loss function $d(z) = 0,\, |z| <= \epsilon,\; d(z) = C(|z| - \epsilon),\, |z| > \epsilon$, has,

$$d^*(\lambda) = \begin{cases} \delta\lambda & |\lambda| \leq C \\ \infty & |\lambda| > C \end{cases}$$

3. The ϵ-insensitive quadratic loss function $d(z) = 0,\, |z| <= \epsilon,\, d(z) = \frac{1}{2\sigma^2}(|z| - \epsilon)^2,\, |z| > \epsilon$, has,

$$d^*(\lambda) = \frac{1}{2}\sigma^2\lambda^2 + \delta\,|\lambda| \quad \lambda \in \mathbb{R}$$

   For this function, range $\partial f = \mathrm{dom} f^* = \mathbb{R}$, and (6) can be optimized for λ using a relatively simple iterative reweighted least squares algorithm. As in the similar classification example, this function is less robust to outliers than the asymptotically linear functions.

4. In general, if $d(z)$ is symmetric with $d(0) = 0$, and $\tilde{d}(z)$ is defined by,

$$\tilde{d}(z) = \begin{cases} 0 & |z| \leq \epsilon \\ d(|z| - \epsilon) & |z| > \epsilon \end{cases}$$

   then the conjugate is given by,

$$\tilde{d^*}(\lambda) = d^*(\lambda) + \epsilon|\lambda|$$

5. The loss function $d(z)$ corresponding to the negative logarithm of the logistic derivate, given by $d(z) = -\log s - \log(1-s) - \log 4$, where $s = (1+\exp(-z))^{-1}$, has,

$$d^*(\lambda) = \begin{cases} -\log(1-|\lambda|) & |\lambda| \leq 1 \\ \infty & |\lambda| > 1 \end{cases}$$

   Since $d(z)$ is asymptotically linear, it is robust to outliers.
6. Huber's loss function, $d(z) = \frac{1}{2}z^2$, $|z| \leq c$, $d(z) = c|z| - c^2/2$, $|z| > c$, has,

$$d^*(\lambda) = \begin{cases} \frac{1}{2}\lambda^2 & |\lambda| \leq c \\ \infty & |\lambda| > c \end{cases}$$

   This function is also robust to outliers.

Curvature of conjugate objective functions. We have seen that linear regression (and thus kernel-based nonlinear regression) using a convex symmetric loss function $d(z)$, is equivalent to linearly constrained minimization of a quadratic form plus the convex conjugate function $d^*(\lambda)$. The two nonquadratic functions d and d^* generally require different optimization approaches, due to the contrasting curvatures of conjugate functions.

The first indication of the curvature relationship between conjugate functions is given by the fact that the Hessians of d and d^* are inverses when they are full rank. Specifically, if $\mathbf{H}_f(\mathbf{z})$ is the Hessian of f at $\mathbf{z}$, then

$$\mathbf{H}_f(\mathbf{z})^{-1} \;=\; \mathbf{H}_{f^*}\big(\phi^*(\mathbf{z})\big) \;=\; \mathbf{H}_{f^*}\big(\nabla f(\mathbf{z})\big)$$

where $\phi^*(\mathbf{z}) = \arg\max_\phi \mathbf{x}^T\phi - f^*(\phi)$ is attained at $\phi^* = \nabla f(\mathbf{z})$ for twice differentiable f and f^*. Here we shall need only the one dimensional case, $f'' = 1/f^{*\prime\prime}$.

A stronger relationship exists, however, for which we require the following concept. For simplicity we limit consideration to the one-dimensional case, which applies directly to the case of separable f on $\mathbb{R}^n$. We define f to be (strictly) *convex with respect to* x^2, or *convex in* x^2, or *square-convex*, on an interval I, if $f(x) = g(x^2)$ with g (strictly) convex and increasing on I^2. A function is (strictly) *square-concave*, etc., on I, if $f(x) = g(x^2)$ with g (strictly) concave and increasing on I^2.

The significance of this definition lies in following inequality, which holds for functions f that are square-concave on (a,b),

$$f(y) - f(x) \leq \phi \cdot \big(y^2 - x^2\big) \qquad \phi \in \partial g(x),\;\; y \in (a,b)$$

where g is defined by $f(x) = g(x^2)$. For g differentiable at x, $\phi = f'(x)/2x$, and,

$$f(y) - f(x) \leq \frac{1}{2}\frac{f'(x)}{x}\big(y^2 - x^2\big) \quad \forall\; y \in (a,b) \tag{12}$$

This inequality can be used to formulate a globally convergent IRLS algorithm for separable, component-wise square-concave functions. Adding the inequalities (12) for the components of $\mathbf{y}$ and $\mathbf{x}$, we have,

$$f(\mathbf{y}) - f(\mathbf{x}) \;\leq\; \frac{1}{2}\,\mathbf{y}^T\boldsymbol{\Pi}(\mathbf{x})\,\mathbf{y} \;-\; \frac{1}{2}\,\mathbf{x}^T\boldsymbol{\Pi}(\mathbf{x})\,\mathbf{x}$$

where $\boldsymbol{\Pi}(\mathbf{x})$ is diagonal with $\big[\boldsymbol{\Pi}(\mathbf{x})\big]_{i,i} \equiv f'(x_i)/x_i$. Thus if we are minimizing a (component-wise) square-concave function f over a convex set C, then the iterative algorithm given by,

$$\mathbf{x}_{k+1} \;\leftarrow\; \arg\min_{\mathbf{x}\in C}\, \mathbf{x}^T\boldsymbol{\Pi}(\mathbf{x}_k)\mathbf{x} \tag{13}$$

is guaranteed to decrease f. Note that (12) also guarantees descent for symmetric, concave functions increasing on $(0,\infty)$, which can lead to increased sparsity in the solution, at the cost of introducing multiple local optima.

Returning to the curvature relationship between conjugate functions, we have the following theorem [5],

Theorem 1 *f is convex and strictly square-concave on the non-negative interval I if and only if f^* is strictly square-convex on I.*

Proof. Suppose for simplicity that f and f^* are differentiable on I. [1] Then f is strictly square-concave if and only if $f'(x)/x$ is strictly decreasing, i.e. if and only if $x < y$ implies,

$$\frac{f'(x)}{x} < \frac{f'(y)}{y} \tag{14}$$

Let $\phi_x = f'(x)$ and $\phi_y = f'(y)$. Then $\phi_x < \phi_y$ since f is strictly convex, and from (17), $x = f^{*\prime}(\phi_x)$ and $y = f^{*\prime}(\phi_y)$. Substituting these into (14), we get,

$$\frac{\phi_x}{f^{*\prime}(\phi_x)} = \frac{f'(x)}{x} < \frac{f'(y)}{y} = \frac{\phi_y}{f^{*\prime}(\phi_y)}$$

or, $f^{*\prime}(\phi_x)/\phi_x > f^{*\prime}(\phi_y)/\phi_y$, which implies that $f^*(\phi)/\phi$ is strictly increasing and f^* is strictly square-convex.

Thus if the primal objective is strictly square-concave, then the dual objective is strictly square-convex, and vice-versa.

Qualitatively, square-concave functions have "less curvature" than the quadratic. These functions correspond to the negative logarithm of super-gaussian densities [6], which have a sharp peak at zero and "heavy tails". Correspondingly, square-concave functions have possibly infinite curvature at zero, and are asymptotically order x^2. The sharp curvature at zero tends to promote sparsity, but it should be observed that if the curvature is infinite at zero, e.g. for $x^{3/2}$, then Newton's method will not work, since it requires twice-differentiability at the solution and square-concave functions generally tend to produce solutions

[1] Differentiability is in fact necessary for square-concavity [1].

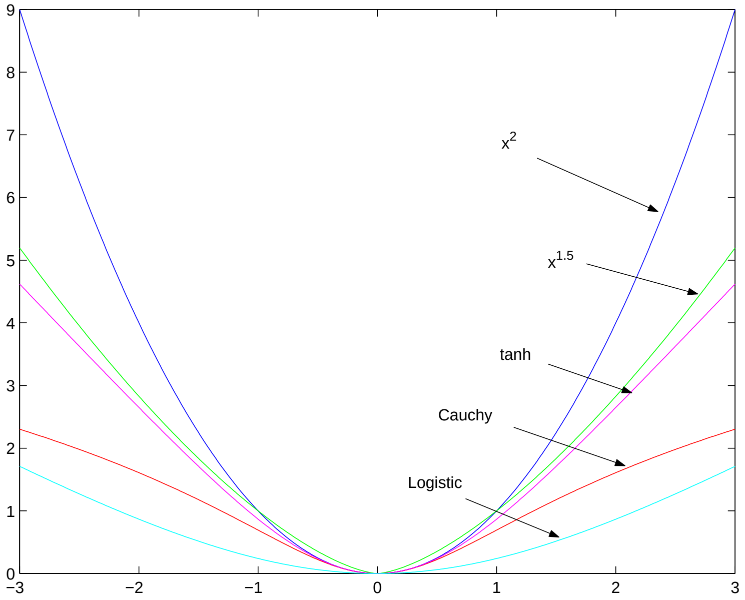

Fig. 3. Some square-concave loss functions corresponding to super-gaussian densities.

with elements close to zero. To optimize these functions we have two options: either use the descent property for global convergence on the primal problem, or use Newton's method on the dual square-convex function, which will generally be twice-differentiable when the primal objective is not.

Square-convex functions qualitatively correspond to sub-gaussian densities, being flat around the origin and rising sharply after crossing unity. The correspondence between square-concavity and square-convexity of the dual is exemplified by the functions $f(x) = (1/p)\, x^p$ and $f^*(\phi) = (1/q)\, \phi^q$, where $1/p + 1/q = 1$. If $p < 2$, then $q > 2$. $f(x)$ is non-twice-differentiable at the origin and tends asymptotically toward linearity (though we still have range $\partial f = \mathbb{R}$), while $f^*(\phi)$ has zero curvature at zero, and increases sharply for $|\phi| > 1$. It is of course not necessary that square-concave functions have infinite curvature at zero, as exemplified by the negative log logistic density.

General algorithms for kernel regression. We restate here the dual of the general regression problem for convenience,

$$\max_{\lambda \in \partial f} \max_{\mu \geq 0} \quad \frac{1}{4\mu}\, \lambda^T \mathbf{K}\, \lambda \, - \, \mathbf{y}^T \lambda \, + \, d^*(\lambda) \, + \, A^2 \mu \quad s.t. \quad \mathbf{e}^T \lambda = 0 \tag{15}$$

From the preceding section, when $d(z)$ is square-convex, $d^*(\lambda)$ is square-concave and the following algorithm can be used to minimize (15). Given μ_{k-1}, let $\mathbf{Q}_k =$

$\frac{1}{2\mu_{k-1}}\mathbf{K} + \mathbf{\Pi}(\lambda_{k-1})$, and set,

$$\lambda_k \;\leftarrow\; \arg\min_{\lambda\in\partial f} \; \frac{1}{2}\,\lambda^T\mathbf{Q}_k\,\lambda + \mathbf{y}^T\lambda \quad s.t. \quad \mathbf{e}^T\lambda = 0$$

and $\mu_k \leftarrow \frac{1}{2}\left\|\lambda_k\right\|_{\mathbf{K}}$, as in the regression case. When range $\partial f = \mathbb{R}$, the algorithm reduces to,

$$\lambda_k = \left(\mathbf{Q}_k^{-1} - \frac{\mathbf{Q}_k^{-1}\mathbf{e}\mathbf{e}^T\mathbf{Q}_k^{-1}}{\mathbf{e}^T\mathbf{Q}_k^{-1}\mathbf{e}}\right)\mathbf{y} \;\;\equiv\;\; \mathbf{r} - \frac{\mathbf{e}^T\mathbf{r}}{\mathbf{e}^T\mathbf{p}}\,\mathbf{r}$$

where $\mathbf{r}$ and $\mathbf{p}$ are defined by the equations $\mathbf{Q}_k\mathbf{r} = \mathbf{y}$ and $\mathbf{Q}_k\mathbf{p} = \mathbf{e}$ respectively.

For robust regression (in the primal space), however, we want $d(z)$ to square-concave, so that $d^*(\lambda)$ is square-convex. In this case we can update λ with a standard Newton step, replacing $\mathbf{\Pi}(\lambda_{k-1})$ by $\mathbf{H}_{f^*}(\lambda_{k-1})$, however we must in principle impose safeguards in the optimization to ensure decrease of the objective.

3.1 Sparsity in kernel methods

An issue with algorithms formulated as optimization problems in Hilbert space is that that they may make liberal use of the training examples in the definition of the optimal hyperplane, even with Vapnik's generalized optimal hyperplane, and ϵ-insensitive loss function. The Relevance Vector Machine [10] attempts to deal with this problem by imposing a prior function on the Lagrange multiplier vector directly and using automatic relevance detection [4] to hone the solution to minimal support in the sample data. Similarly, [2] uses an algorithm for regression that is equivalent to the FOCUSS algorithm of [7].

These algorithms treat the regression problem as one of estimation of λ in the linear model $\mathbf{K}\lambda = \mathbf{y}$, with a sparsity inducing prior on λ, thus essentially bypassing the consideration of Hilbert space and working directly in the dual space. It might be wondered what problem these dual space algorithms are solving in the primal space. As noted, the dual problem can only be formulated when the loss function $d(z)$ is convex (in the regression case), in which case the conjugate function $d^*(\lambda)$ is also convex. In order to achieve a high degree of sparsity (concentration) in the solution, the algorithms mentioned, in effect, impose a *concave* loss function on λ. Since there is no dual problem corresponding to minimization of a concave function, there is no corresponding optimization problem being solved in Hilbert space. Nevertheless, the algorithms of [10, 2] exhibit very little if any performance degradation compared the “legitimate” kernel methods.

Support vector machines are themselves said to exhibit sparsity in the solution, and this sparsity is presented as contributing to generalization ability [11, §10.3]. Unfortunately, however, the number of support vectors generally scales linearly with the number of training samples. Furthermore, while bounds can be derived on generalization based on number of support vectors, in practice, the real control on generalization is the structural risk minimization parameter h, which controls the VC dimension of the function domain used employed for optimization. Generally, having more support vectors contributes to generalization ability by improving the "optimality" of the optimal hyperplane at the given function capacity of the structure element, and sacrificing support vectors generally leads to degraded performance.

Appendix: Convex analysis

We follow [8] in considering functions defined on all of $\mathbb{R}^n$, possibly taking on the values $+\infty$ and $-\infty$. A function function $f : \mathbb{R}^n \to \mathbb{R}^m$ is *convex* if $f\big(\alpha x + (1-\alpha)y\big) \leq \alpha f(x) + (1-\alpha)f(y)$, $0 \leq \alpha \leq 1$, and *proper* if it is nowhere $-\infty$ and somewhere finite. A function is *concave* and proper if $-f$ is convex and proper. The *domain* of f, denoted $\mathrm{dom} f$, is the set on which f is finite, and f is *closed* if $\mathrm{dom} f$ is closed. Here we shall consider only closed and proper convex and concave functions.

Conjugacy and subdifferentials. If f is convex, then the *convex conjugate* of f, denoted f^*, is defined by,

$$f^*(\phi) = \sup_x \phi^T x - f(x) \tag{16}$$

If f is concave, then the *concave conjugate* of f, also denoted f^*, is defined by $f^*(\phi) = \inf_x \phi^T x - f(x)$.

For convex f, the *subdifferential of f at x* [8, pp. 214-5], denoted $\partial f(x)$, is the set of $z \in \mathbb{R}^n$ such that,

$$f(y) - f(x) \geq z^T(y - x) \quad \forall\, y \in \mathbb{R}^n$$

For f concave, $\partial f(x)$ is similarly defined to be the set of $z \in \mathbb{R}^n$ such that $f(y) - f(x) \leq z^T(y-x)\ \forall\, y \in \mathbb{R}^n$. The set-valued mapping $\partial f : x \to \partial f(x)$ is called the *subdifferential* of f. If $\partial f(x)$ is not empty, f is *subdifferentiable* at x. If f is differentiable at x, then $\partial f(x)$ is a singleton containing the gradient, $\nabla f(x)$.

Subdifferentials and conjugacy are closely related by the following fact [8, Cor. 23.5.1], which holds for closed, proper convex and concave functions,

$$\phi \in \partial f(x) \quad \text{if and only if} \quad x \in \partial f^*(\phi) \tag{17}$$

which can be written $\partial f^{-1} = \partial f^*$ in the sense of set-valued mappings. We thus have,

$$\mathrm{dom} f = \mathrm{range}\, \partial f^*\,, \quad \mathrm{dom} f^* = \mathrm{range}\, \partial f \tag{18}$$

By an abuse of notation, we shall denote $x \in \mathrm{range}\, \partial f$ simply by $x \in \partial f$.

Convex programs and Lagrangians. A *convex program* (P) consists of a convex objective function $f(x)$, a set of m linear equality constraints represented by the equation $Ax = b$, $A \in \mathbb{R}^{m \times n}$, and a set of r convex inequality constraints, $g_1(x) \leq 0, \ldots, g_r(x) \leq 0$.[2] The *primal* problem is to minimize $f(x)$ subject to

[2] In [8], a convex constraint set $C \subset \mathbb{R}^n$ is included in definitions and theorems concerning convex programming, noting the alternative of defining $f(x) = \infty$ outside C, which we follow here.

$Ax = b$, $g_1(x) \leq 0, \ldots, g_r(x) \leq 0$. The *Lagrangian* of (P) is defined by,

$$L(x, \lambda, \mu) = \begin{cases} f(x) + \lambda^T(Ax - b) + \sum_{i=1}^{r} \mu_i \, g_i(x) & \mu_i \geq 0 \; \forall \, i \\ -\infty & \exists \, i \;\; s.t. \;\; \mu_i < 0 \end{cases}$$

The primal problem is trivially equivalent to,

$$\inf_x f(x) \; = \; \inf_x \, \sup_\lambda \, \sup_{\mu \geq 0} \, L(x, \lambda, \mu) \tag{19}$$

where $\mu \geq 0$ denotes element-wise non-negativity of $\mu \in \mathbb{R}^r$. The following theorem [8, Thm. 28.4, Cor. 28.4.1] gives sufficient conditions for the interchange of inf and sup and the formulation of the *dual* problem.

Theorem 2 *Given a convex program* (P) *in which* $\inf_x f(x) > -\infty$*, and* $\exists \, x \in \mathbb{R}^n$ *such that* $g_i(x) < 0$*,* $i = 1, \ldots, r$*, i.e. such that each of the* r *inequality constraints is strictly satisfied (the* Slater condition*), then,*

$$\inf_{x \in C} f(x) \; = \; \inf_x \, \sup_\lambda \, \sup_{\mu \geq 0} \, L(x, \lambda, \mu) \; = \; \sup_\lambda \, \sup_{\mu \geq 0} \, \inf_x \, L(x, \lambda, \mu)$$

Thus, under the conditions of the theorem, minimization of f over $x \in \mathbb{R}^n$ (the primal problem) is equivalent to maximization of $\inf_x \, L(x, \lambda, \mu)$ over $\lambda \in \mathbb{R}^m$, $\mu \in \mathbb{R}^r$, $\mu \geq 0$ (the dual problem). The maximization of a concave function subject to linear equality and convex inequality constraints is referred to as a *concave program.*

References

1. E. F. Beckenbach. Generalized convex functions. *Bull. Am. Math. Soc.*, 43:363–371, 1937.
2. M. Figueiredo. Adaptive sparseness using Jeffreys prior. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, Cambridge, MA, 2002. MIT Press.
3. T. S. Jaakkola and D. Haussler. Probabilistic kernel regression models. In David Heckerman and Joe Whittaker, editors, *Proceedings of the 7th Intl. Work. on AI and Statistics*. Morgan Kaufmann Publishers, Inc., San Francisco, CA, 1999.
4. D. J. C. Mackay. Comparison of approximate methods for handling hyperparameters. *Neural Computation*, 11(5):1035–1068, 1999.
5. J. Palmer and K. Kreutz-Delgado. A globally convergent algorithm for maximum likelihood estimation with non-gaussian priors. In *Proceedings of the 36th Asilomar Conference on Signals and Systems*. IEEE, 2002.
6. J. Palmer and K. Kreutz-Delgado. A general framework for component estimation. In *Proceedings of the 4th International Symposium on Independent Component Analysis*, 2003.
7. B. D. Rao and I. F. Gorodnitsky. Sparse signal reconstruction from limited data using FOCUSS: a re-weighted minimum norm algorithm. *IEEE Trans. Signal Processing*, 45:600–616, 1997.
8. R. T. Rockafellar. *Convex Analysis.* Princeton, 1970.

9. B. Schölkopf, A. Smola, R. Williamson, and P. Bartlett. New support vector algorithms. *Neural Computation*, 12:1207–1245, 2000.
10. M. E. Tipping. Sparse Bayesian learning and the Relevance Vector Machine. *Journal of Machine Learning Research*, 1:211–244, 2001.
11. V. Vapnik. *Statistical Learning Theory.* John Wiley & Sons, Inc., 1998.