The Coherent Loss Function for Classification

Wenzhuo Yang A0096049@NUS.EDU.SG
Department of Mechanical Engineering, National University of Singapore, Singapore 117576

Melvyn Sim DSCSIMM@NUS.EDU.SG
Department of Decision Sciences, National University of Singapore, Singapore 117576

Huan Xu MPEXUH@NUS.EDU.SG
Department of Mechanical Engineering, National University of Singapore, Singapore 117576

Abstract

A prediction rule in binary classification that aims to achieve the lowest probability of misclassification involves minimizing over a non-convex, 0-1 loss function, which is typically a computationally intractable optimization problem. To address the intractability, previous methods consider minimizing the *cumulative loss* – the sum of convex surrogates of the 0-1 loss of each sample. We revisit this paradigm and develop instead an *axiomatic* framework by proposing a set of salient properties on functions for binary classification and then propose the *coherent loss* approach, which is a tractable upper-bound of the empirical classification error over the *entire* sample set. We show that the proposed approach yields a strictly tighter approximation to the empirical classification error than any convex cumulative loss approach while preserving the convexity of the underlying optimization problem, and this approach for binary classification also has a robustness interpretation which builds a connection to robust SVMs.

1. Introduction

The goal of supervised learning is to predict an unobserved output value y from an observed input $\mathbf{x}$. This is achieved by learning a function relationship $y \approx f(\mathbf{x})$ from a set of observed training examples $\{(y_i, \mathbf{x}_i)\}_{i=1}^m$. The quality of predictor $f(\cdot)$ is often measured by some loss function $\ell(f(\mathbf{x}),\, y)$. A typical statistical setup in machine learning assumes that all training data and testing samples are IID samples drawn from an unknown distribution μ, and the goal is to find a predictor $f(\cdot)$ such that the expected loss $\mathbb{E}_{(y,\mathbf{x})\sim\mu}\ell(f(\mathbf{x}),\, y)$ is minimized. Since μ is unknown, the expected loss is often replaced by the empirical loss

$$L_{emp}(f) \triangleq \frac{1}{m}\sum_{i=1}^{m} \ell(f(\mathbf{x}_i),\, y_i). \tag{1}$$

Minimizing $L_{emp}(f)$, as well as numerous regularization based variants of it, is one of the fundamental cornerstones of statistical machine learning (e.g., Vapnik & Lerner, 1963; Vapnik & Chervonenkis, 1991; Poggio et al., 2004).

This paper focuses on binary classification problems, where $y \in \{-1, +1\}$. A point $(y, \mathbf{x})$ is correctly predicted if $\operatorname{sign}(f(\mathbf{x})) = y$, and its classification error is given by the 0-1 loss $\ell(f(\mathbf{x}_i),\, y_i) = \mathbf{1}(y \neq \operatorname{sign}(f(\mathbf{x})) = \mathbf{1}(yf(\mathbf{x}) \leq 0)$. Due to the non-convexity of the indicator function, minimizing the empirical classification error $\sum_i \mathbf{1}(y_i f(\mathbf{x}_i) \leq 0)$ is known to be NP-hard even to approximate (Arora et al., 1997; Ben-David et al., 2003). A number of methods have been proposed to mitigate this computational difficulty, all based on the idea that to minimize the "cumulative loss", which is the sum of individual losses given by,

$$L_\phi(f) \triangleq \frac{1}{m}\sum_{i=1}^{m} \phi(yf(\mathbf{x}))$$

where $\phi(\cdot)$ is a convex upper bound of the classification error $\mathbf{1}(yf(\mathbf{x}) \leq 0)$. For example, AdaBoost (Freund & Schapire, 1997; Friedman et al., 2000; Schapire & Singer, 1999) employs the exponential loss function $\exp(-yf(\mathbf{x}))$, and Support Vector Machines (SVMs) (Boser et al., 1992; Cortes & Vapnik, 1995) employ a hinge-loss function $\max\{1 - yf(\mathbf{x}), 0\}$.

In this paper we revisit this paradigm, and introduce a notion termed *coherent loss*, as opposed to cumulative loss

Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

used in the conventional approach. Briefly speaking, instead of using an upper bound of the *individual* classification error (the 0-1 loss), we propose to use a tractable upper bound of the *total* empirical classification error for the whole training set. That is, we look for $\Phi : \Re^m \mapsto \Re$ such that

$$\Phi(c_1, \cdots, c_m) \geq \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}(c_i \leq 0), \ \forall (c_1, \cdots, c_m) \in \Re^m.$$

Intuitively, since coherent loss functions are more general than cumulative loss functions, one may expect to obtain a tighter and still tractable bound of the empirical classification error via coherent loss function. We formalize this intuition in this paper. Specifically, our contributions include the followings.

In Section 2, we consider a principled approach by formalizing the salient properties of functions, termed as *coherent classification loss* functions, that could be used to quantify the performance of a classification rule. These functions have dual-representations which enable us to identify the *minimal coherent classification loss* function, which, loosely speaking, is the coherent classification loss function that best approximates the 0-1 loss, which also achieves a tighter bound of the empirical classification error than any convex cumulative loss. We show that optimizing this function is equivalent to a convex optimization problem, and hence tractable.

In Section 3, we consider an equivalent form of the coherent loss function and then provide several applications of this loss function in classification problems. We remark that a tighter approximation of the 0-1 loss can potentially reduce the impact of outliers on the classification accuracy. Cumulative loss function may significantly deviate from the 0-1 loss when $c \ll 0$. Consequently, a misclassified outlier can incur a huge loss, and prevents an otherwise perfect prediction rule from being selected. This sensitivity can be mitigated by a tighter approximation.

Section 4 provides a statistical interpretation of minimizing the coherent loss function. Section 5 reports the experimental results which show that our classification method outperforms the standard SVM when additional constraints are imposed on the decision function.

Notations: We use boldface letters to represent column vectors, and capital letters for matrices. We reserve $\mathbf{e}$ for special vectors: $\mathbf{e}_i$ is the vector whose i-th entry is 1, and the rest are 0; $\mathbf{e}_N$, where N is an index set, is the vector that for all $i \in N$, the corresponding entry equals 1, and zero otherwise; $\mathbf{e}$ is the vector with all entries equal to 1. The i-th entry of a vector $\mathbf{x}$ is denoted by x_i. We use $[c]_+$ to denote $\max\{0, c\}$ and $\mathbf{1}[\cdot]$ to denote the indicator function, and let $\mathcal{P}_m$ be the set of all $m \times m$ permutation matrices and $\mathbf{I}_p$ be the $p \times p$ identity matrix.

2. Coherent Classification Loss Function

We now propose the notion of coherent classification loss functions based on an axiomatic approach. Along the way, we show the existence of a "tight" coherent classification loss function which can achieve better approximation of the empirical classification error than *any* convex cumulative loss. The definition of the coherent classification loss function is motivated from analyzing the salient properties of functions used to quantify the performance of a classification rule. A natural approach is to elicit these properties from the *classification error*. Specifically, given $u_1, \cdots, u_m$ where u_i the "decision value" of the ith sample, e.g. $u_i = y_i f(\mathbf{x}_i)$, the classification error $\varrho : \Re^m \mapsto [0, 1]$ is given by

$$\varrho(u_1, \ldots, u_m) = \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}[u_i < 0]. \tag{2}$$

We will next propose a set of properties and that functions endowed with these properties are known as coherent classification loss functions.

2.1. Salient Properties and Representation Theorem

We elicit the five salient properties from the classification error as follows. Consider $\rho(\cdot) : \Re^m \to [0, 1]$.

Property 1 (Complete classification)**.** $\rho(\mathbf{u}) = 0$ *if and only if* $\mathbf{u} \geq \mathbf{0}$.

Complete classification essentially says if every sample is correctly classified, then it is optimal.

Property 2 (Misclassification avoidance)**.** *If* $\mathbf{u} < \mathbf{0}$*, then* $\rho(\mathbf{u}) = 1$.

This property states that if all samples are misclassified, then it is the worst classification and hence $\rho(\cdot)$ achieves the maximal value.

Property 3 (Monotonicity)**.** *If* $\mathbf{u}_1 \geq \mathbf{u}_2$ *then* $\rho(\mathbf{u}_1) \leq \rho(\mathbf{u}_2)$.

Monotonicity requires that if a decision better classifies every sample, then it is more desirable.

Property 4 (Order invariance)**.** *For any* $P \in \mathcal{P}_m$*, we have* $\rho(\mathbf{u}) = \rho(P\mathbf{u})$.

Order invariance essentially states that the order of the samples does not matter. This is natural in the classification problem, since each sample is drawn IID, and is treated equally.

Property 5 (Scale invariance)**.** *For all* $\alpha > 0$*,* $\rho(\alpha\mathbf{u}) = \rho(\mathbf{u})$.

Scale invariance is a property that the classification error function satisfies. It essentially means that changing the

scale does not affect the preference between classifiers. While it may be debatable whether scale invariance is as necessary as other properties, indeed as we show later in this section, this property can be relaxed.

Definition 1 (Coherent Classification Loss)**.** *A function $\rho(\cdot) : \Re^m \to [0,1]$ is a* coherent classification loss function *(CCLF) if it satisfies Property 1 to 5, and is quasi-convex and lower semi-continuous.*

Here, quasi-convexity and semi-continuity are introduced to for tractability. Our first result is a (dual) representation theorem of any CCLF. We need the following definition first.

Definition 2 (Admissible Class)**.** *A class of sets $\mathbf{V}_k \subseteq \Re^m$ parameterized by $k \in [0,1]$ is called* admissible class*, if they satisfy the following properties:*

1. *For any $k \in [0,1]$, $\mathbf{V}_k$ is a closed, convex cone, and is order invariant. Here, being order invariant means that $\mathbf{v} \in \mathbf{V}$ implies $P\mathbf{v} \in \mathbf{V}$ for any $P \in \mathcal{P}_m$;*
2. *$k \le k'$ implies $\mathbf{V}_k \subseteq \mathbf{V}_{k'}$;*
3. *$\mathbf{V}_1 = \mathrm{cl}(\lim_{k\uparrow 1} \mathbf{V}_k)$ and $\mathbf{V}_0 = \lim_{k\downarrow 0} \mathbf{V}_k$.*
4. *$\mathbf{V}_1 = \Re^m_+$;*
5. *For any $\lambda > 0$, we have $\lambda \mathbf{e} \in \mathbf{V}_0$.*

Theorem 1 (Representation Theorem)**.** *A function $\rho(\cdot)$ is a CCLF if and only if it can be written as*

$$\rho(\mathbf{u}) = 1 - \sup\{k \in [0,1] | \sup_{\mathbf{v} \in \mathbf{V}_k} (-\mathbf{v}^\top \mathbf{u}) \le 0\}, \quad (3)$$

for an admissible class $\{\mathbf{V}_k\}$. Here sup *over an empty set is set as* 0*.*

Proof. We sketch the proof and leave the details in the supplementary material. The "if" part is relatively easy, by checking that any function $\rho(\mathbf{u}) = 1 - \sup\{k \in [0,1] | \sup_{\mathbf{v} \in \mathbf{V}_k} (-\mathbf{v}^\top \mathbf{u}) \le 0\}$ for some admissible class $\{\mathbf{V}_k\}$ satisfies all properties required for a CCLF.

The "only if" part requires more work. We want to show that given a function $\rho(\cdot)$ which is a CCLF, it can be represented as (3) for some admissible class $\{\mathbf{V}_k\}$. The proof consists of three steps: We first show that $\rho(\cdot)$ can be represented as $\rho(\mathbf{u}) = 1 - \sup\{k \in [0,1] | \sup_{\mathbf{v} \in \overline{\mathbf{V}}_k} (-\mathbf{v}^\top \mathbf{u}) \le 0\}$, for some $\{\overline{\mathbf{V}}_k\}$, not necessarily admissible. This essentially follows from a result in Brown & Sim (2009). We then show that we can replace $\overline{\mathbf{V}}_k$ by a class of closed, convex, order-invariant, cones $\mathbf{V}_k$. Specifically, we can pick $\mathbf{V}_k \triangleq \mathrm{cl}(\mathrm{cc}(\mathrm{or}(\overline{\mathbf{V}}_k)))$, where $\mathrm{or}(\cdot)$ (respectively $\mathrm{cc}(\cdot)$) is the minimal **or**der invariant (respectively, **c**onvex **c**one) superset. Finally we show that $\{\mathbf{V}_k\}$ is admissible, by checking that all properties in Definition 2 are satisfied, to complete the proof. □

2.2. Minimal coherent classification loss function

This section shows that among all CCLF functions that upper-bound the classification error, there exists a minimal (i.e., best) one.

Theorem 2. *Define $\overline{\rho}(\cdot) : \Re^m \mapsto [0,1]$ as follows*

$$\overline{\rho}(\mathbf{u}) = \frac{\max\{t : \sum_{i=1}^t u_{(i)} < 0\}}{m},$$

where $\{u_{(i)}\}$ is a permutation of $\{u_i\}$ in a non-decreasing order, and max *over an empty set is taken as zero. Then the following holds.*

1. *$\overline{\rho}(\cdot)$ is a CCLF, and is an upper-bound of the* classification error*, i.e., $\overline{\rho}(\mathbf{u}) \ge \varrho(\mathbf{u}),\ \forall \mathbf{u} \in \Re^m$.*
2. *Let $\overline{\mathbf{V}}_k \subset \Re^m$ satisfy that if $k = 0$, then $\overline{\mathbf{V}}_k = \mathrm{conv}\{\lambda \mathbf{e} | \lambda > 0\}$; and if $\frac{s}{m} < k \le \frac{s+1}{m}$ for $s = 0, \cdots, m-1$, then*

   $$\overline{\mathbf{V}}_k = \mathrm{conv}\left\{\lambda \mathbf{e}_N | \forall \lambda > 0, \forall N : |N| = m - s\right\},$$

   where N is an index set. Then $\{\overline{\mathbf{V}}_k\}$ is an admissible class corresponding to $\overline{\rho}(\cdot)$.
3. *$\overline{\rho}(\cdot)$ is the* tightest *CCLF bound. That is, if $\rho'(\cdot)$ is a CCLF function and satisfies $\rho'(\mathbf{u}) \ge \varrho(\mathbf{u})$ for all $\mathbf{u} \in \Re^m$, then $\rho'(\mathbf{u}) \ge \overline{\rho}(\mathbf{u})$ for all $\mathbf{u} \in \Re^m$.*

Proof. We provide a sketch of the proof and leave the details to the supplementary material. Claim 1 is relatively straightforward. It is also easy to see that $\overline{\mathbf{V}}_k$ is an admissible set. So one only needs to show that $\overline{\mathbf{V}}_k$ is the set corresponding to $\overline{\rho}(\cdot)$, to establish Claim 2. To show Claim 3, we let $\{\mathbf{V}'_k\}$ be an admissible class corresponding to $\rho'(\cdot)$, and show that $\lambda \mathbf{e}_N \in \mathbf{V}'_k$, which further implies $\overline{\mathbf{V}}_k \subseteq \mathbf{V}'_k$. This establishes claim 3. □

We next show that *scale invariance* can be relaxed. Indeed, for any quasi-convex upper bound of classification error that satisfies other properties, the minimal CCLF is a tighter bound.

Theorem 3. *Let $\hat{\rho} : \Re^m \mapsto [0,1]$ be a quasi-convex function that satisfies complete classification, misclassification avoidance, monotonicity, order invariance, and that $\hat{\rho}(\mathbf{u}) \ge \varrho(\mathbf{u})$. Then there exists a CCLF $\rho(\cdot)$ such that*

$$\varrho(\mathbf{u}) \le \overline{\rho}(\mathbf{u}) \le \rho(\mathbf{u}) \le \hat{\rho}(\mathbf{u}), \quad \forall \mathbf{u} \in \Re^m.$$

Proof. We sketch the proof. The main idea is to construct such a function

$$\rho(\mathbf{u}) \triangleq \lim_{\epsilon \downarrow 0} [\min_{\gamma > 0} \hat{\rho}((\mathbf{u} + \epsilon)/\gamma)],$$

and show that $\rho(\cdot)$ is a CCLF and $\rho(\hat{\mathbf{u}}) \ge \rho(\mathbf{u}) \ge \varrho(\mathbf{u})$. Finally, since $\overline{\rho}(\cdot)$ is the minimal CCLF, this completes the proof. □

One important property of $\overline{\rho}(\cdot)$ is that it achieves better approximation of the empirical classification error than any *convex cumulative loss*.

Theorem 4. *If $f(\cdot)$ is a convex function and an upper bound of the 0-1 loss function, then for any $\mathbf{u} = (u_1, \cdots, u_m)$, we have $\varrho(\mathbf{u}) \le \overline{\rho}(\mathbf{u}) \le \frac{1}{m}\sum_{i=1}^m f(u_i)$.*

Proof. Without loss of generality, assume $(u_1, \cdots, u_m)$ are in a non-decreasing order. Let $p \triangleq \max\{i | u_i < 0\}$ and $q \triangleq \max\{t | \sum_{i=1}^t u_i < 0\}$, then $\sum_{i=1}^q u_i = \sum_{i=1}^p u_i + \sum_{i=p+1}^q u_i < 0$.

Since $f(\cdot)$ is convex and $f(x) \ge \mathbf{1}[x \le 0]$, there exists $k \le 0$ such that $f(x) \ge \max\{kx+1, 0\}$ (this can be done for example by taking k as a subgradient of $f(x)$ at $x = 0$). If $k = 0$, then $\frac{1}{m}\sum_{i=1}^m f(u_i) \ge 1 \ge \overline{\rho}(\mathbf{u})$, the theorem holds. Otherwise $k < 0$, we have

$$\begin{aligned}\sum_{i=1}^m f(u_i) &\ge \sum_{i=1}^p (ku_i + 1) + \sum_{i=p+1}^m f(u_i)\\ &= p + k\sum_{i=1}^p u_i + \sum_{i=p+1}^m f(u_i)\\ &> p - k\sum_{i=p+1}^q u_i + \sum_{i=p+1}^m f(u_i)\\ &\ge p + \sum_{i=p+1}^q (f(u_i) - ku_i).\end{aligned}$$

Note that $u_i \ge 0$ for $i = p+1, \cdots, m$, then if $u_i \ge -\frac{1}{k}$, $f(u_i) - ku_i \ge -ku_i \ge 1$. Otherwise $f(u_i) - ku_i \ge ku_i + 1 - ku_i = 1$. Hence, $p + \sum_{i=p+1}^q (f(u_i) - ku_i) \ge p + (q-p) = q$. By the definition of $\overline{\rho}(\mathbf{u})$, the theorem holds. □

2.3. Optimization with the coherent loss function

We now discuss the computational issue of optimization of the minimal CCLF $\overline{\rho}(\cdot)$. Indeed, we show that this can be converted to a tractable convex optimization problem. Specifically, we consider the following problem on variables $(\mathbf{u}, \mathbf{w})$:

$$\begin{aligned}\min \quad & \overline{\rho}(\mathbf{u})\\ \text{s.t.} \quad & f_j(\mathbf{u}, \mathbf{w}) \le 0; \quad j = 1, \cdots, n,\end{aligned} \tag{4}$$

where $f_i(\cdot, \cdot)$ are convex functions. We have the following theorem.

Theorem 5. *Assume complete classification is not achievable, i.e., there is no feasible $(\mathbf{u}, \mathbf{w})$ with $\mathbf{u} \ge \mathbf{0}$. Let $(\mathbf{s}^*, \mathbf{t}^*, h^*)$ be an optimal solution to the following optimization problem:*

$$\begin{aligned}\min_{h,\mathbf{s},\mathbf{t}} \quad & \frac{1}{m}\sum_{i=1}^m [1 - s_i]_+\\ s.t. \quad & hf_j(\mathbf{s}/h, \mathbf{t}/h) \le 0;\ j = 1, \cdots, n;\\ & h > 0.\end{aligned} \tag{5}$$

Then $(\mathbf{s}^*/h^*, \mathbf{t}^*/h^*)$ is an optimal solution to Problem (4).

Proof. We provide a sketch of the proof. We first show that the level set of Problem (4), i.e., $\mathcal{U}_i \triangleq \{(\mathbf{u}, \mathbf{w}) \mid \overline{\rho}(\mathbf{u}) \le 1 - i/m;\ f_j(\mathbf{u}, \mathbf{w}) \le 0, \forall j\}$ for $i = 1, \cdots, m$, equals the following

$$\{(\mathbf{u}, \mathbf{w}) | \ \exists d : \sum_{i=1}^m [d - u_i]_+ \le (m - i + 1)d; \quad f_j(\mathbf{u}, \mathbf{w}) \le 0,\ \forall j.\}$$

This can be proved by applying the Theorem 2, and then using duality of linear program. This set can further be shown equivalent to the feasible set of

$$\begin{aligned}&\sum_{i=1}^m [1 - u_i/d]_+ \le (m - i + 1)\\ &f_j(\mathbf{u}, \mathbf{w}) \le 0; \quad j = 1, \cdots, n.\end{aligned} \tag{6}$$

Thus, finding the optimal solution to Problem (4) is equivalent to solve the following problem

$$\begin{aligned}\min \quad & \sum_{i=1}^m [1 - u_i/d]_+\\ \text{s.t.} \quad & f_j(\mathbf{u}, \mathbf{w}) \le 0; \quad j = 1, \cdots, n;\\ & d > 0.\end{aligned} \tag{7}$$

Then let $h = 1/d$, $\mathbf{s} = h\mathbf{u}$ and $\mathbf{t} = h\mathbf{w}$, the theorem is established. □

Notice that $hf_j(\mathbf{s}/h, \mathbf{t}/h)$ is the perspective function of $f_j(\cdot, \cdot)$, and is hence jointly convex to $(h, \mathbf{s}, \mathbf{t})$ (Boyd & Vandenberghe, 2004). Thus, Problem (5) is equivalent to a tractable convex optimization problem.

3. Equivalent Formulation and Applications

From Theorem 5, when there is no $(\mathbf{u}, \mathbf{w})$ such that $\mathbf{u} \ge \mathbf{0}$ and $f_j(\mathbf{u}, \mathbf{w}) \le 0$ for $j = 1, \cdots, n$, Problem (4) is equivalent to minimizing the following optimization problem:

$$\begin{aligned}\min \quad & \Phi(\mathbf{u})\\ \text{s.t.} \quad & f_j(\mathbf{u}, \mathbf{w}) \le 0; \quad j = 1, \cdots, n,\end{aligned} \tag{8}$$

where $\Phi(\mathbf{u})$ is defined by

$$\Phi(\mathbf{u}) \triangleq \min_{\gamma > 0} \frac{1}{m}\sum_{i=1}^m [1 - u_i/\gamma]_+. \tag{9}$$

From this formulation, we also show, from another perspective, that minimizing the coherent loss function is equivalent to minimizing a "tighter" upper bound of the 0-1 loss function, or in other words, the coherent loss function achieves better approximation of the empirical classification error than any convex cumulative loss.

Theorem 6. *Let $\phi : \Re \mapsto \Re^+$ be a non-increasing, convex function that satisfies*

$$\phi(c) \geq \mathbf{1}(c \leq 0), \quad \forall c \in \Re.$$

Then we have for all $\mathbf{u} \in \Re^m$:

$$\frac{1}{m}\sum_{i=1}^{m} \mathbf{1}(u_i \leq 0) \leq \Phi(\mathbf{u}) \leq \frac{1}{m}\sum_{i=1}^{m} \phi(u_i).$$

Proof. Recall that the hinge-loss $\phi_1^*(c) \triangleq [1-c]_+$ is the tightest convex bound of 0-1 loss which has a derivative (or sub-gradient) -1 at $c = 0$ (e.g., (Schölkopf & Smola, 2002)). That is, if a convex function $\phi(\cdot)$ satisfies $\phi(c) \geq \mathbf{1}(c \leq 0), \forall c$, and also satisfies $-1 \in \partial\phi(0)$, then $\phi_1^*(c) \leq \phi(c)$ for all c. Similarly, $\phi_\gamma^*(c) \triangleq \max[1 - c/\gamma]_+$ is the tightest convex bound of 0-1 loss with a derivative $-1/\gamma$ at $x = 0$. Since $\phi(\cdot)$ is non-increasing, it can not have positive derivative at $c = 0$. Thus, $\Phi(\cdot)$ is a tighter bound than any non-increasing, convex cumulative loss functions. □

Recall that (8) is equivalent to the following convex optimization problem

$$\begin{aligned} \min_{h,\mathbf{s},\mathbf{t}} \quad & \frac{1}{m}\sum_{i=1}^{m}[1-s_i]_+ \\ \text{s.t.} \quad & hf_j(\mathbf{s}/h, \mathbf{t}/h) \leq 0;\ j = 1, \cdots, n; \\ & h > 0, \end{aligned} \tag{10}$$

which can be solved efficiently. We now provide some applications of the proposed coherent loss function.

At first, we illustrate with an example, that the proposed coherent loss function can be more robust to outliers. Let $\mathbf{u}^1, \mathbf{u}^2 \in \mathbb{R}^{100}$ be the followings: $\mathbf{u}^1 = (-1000, 1000, 1000, \cdots, 1000)$, and $\mathbf{u}^2 = (+1, -1, +1, -1, \cdots, +1, -1)$. In this case, $\mathbf{u}^2$ appears to be a less favorable classification since 50% of samples are misclassified. It is easy to check that $\mathbf{u}^1$ incurs a much larger hinge-loss than $\mathbf{u}^2$, even though only one sample is misclassified. In contrast, the coherent loss of $\mathbf{u}^1$ is no more than 0.02 (take $\gamma = 1/1000$), and that of $\mathbf{u}^2$ is at least 0.5 (since 50% samples are misclassified, and the coherent loss is an upper bound). Thus, the coherent loss is more robust in this example, partly because it better approximates the 0-1 loss, and hence is less affected by large outliers. See Figure 1.

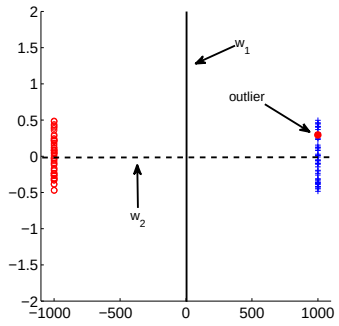

Figure 1. Illustration of the effect of outliers to the cumulative loss vs the coherent loss. Here, $\mathbf{w}_1$ has a margin $\mathbf{u}^1$, and $\mathbf{w}_2$ has a margin $\mathbf{u}^2$. The cumulative loss approach will pick $\mathbf{w}_2$, where the proposed method will pick $\mathbf{w}_1$, which is a better classification.

EXAMPLE: LINEAR SVM

We illustrate the proposed method with the linear classification problem, and in particular, the linear Support Vector Machines algorithm (SVMs) (Boser et al., 1992; Cortes & Vapnik, 1995; Schölkopf & Smola, 2002). Given m training samples $(y_i, \mathbf{x}_i)_{i=1}^m$, the goal is to find a hyperplane that correctly classify as many training samples as possible with a large margin, which leads to the following formulation:

$$\begin{aligned} \min \quad & \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}[y_i(\mathbf{w}^\top \mathbf{x}_i + b) \leq 0] \\ \text{s.t.} \quad & \|\mathbf{w}\|_2 \leq C \end{aligned} \tag{11}$$

for a given $C > 0$. Since the objective function is non-convex, Problem (11) is an intractable problem. Hence, SVM uses the hinge-loss function $\phi_1^*(c) = [1-c]_+$ as a convex surrogate.

Following the proposed coherent loss function approach, we minimize the 0-1 loss function with margin $a \geq 0$: $\frac{1}{m}\sum_{i=1}^{m} \mathbf{1}[y_i(\mathbf{w}^\top \mathbf{x}_i + b) \leq a]$ and replace this objective function by the coherent loss function $\overline{\rho}(\mathbf{u})$ where $u_i = y_i(\mathbf{w}^\top \mathbf{x}_i + b) - a$ (Margin a makes the condition in Theorem 5 hold, and the approximation of this 0-1 loss function by using the hinge-loss function still leads to the standard SVM). Then we obtain the following formulation,

$$\begin{aligned} \min_{\mathbf{w},b,\gamma>0} \quad & \frac{1}{m}\sum_{i=1}^{m}[1 - (y_i(\mathbf{w}^\top \mathbf{x}_i + b) - a)/\gamma]_+ \\ \text{s.t.} \quad & \|\mathbf{w}\|_2 \leq C. \end{aligned} \tag{12}$$

As discussed above, we can change variables $h = 1/\gamma$, $\hat{\mathbf{w}} = \mathbf{w}/\gamma$ and $\hat{b} = b/\gamma$, and simplify Formulation (12) as the following:

$$\begin{aligned} \min_{\hat{\mathbf{w}},\hat{b},h>0} \quad & \frac{1}{m}\sum_{i=1}^{m}[1 + ah - y_i(\hat{\mathbf{w}}^\top \mathbf{x}_i + \hat{b})]_+ \\ \text{s.t.} \quad & \|\hat{\mathbf{w}}\|_2 \leq hC. \end{aligned}$$

This is also equivalent to the robust formulation of SVM (Shivaswamy et al., 2006):

$$\begin{aligned}\min_{\mathbf{w},b,\xi} \quad & \sum_{i=1}^m \xi_i \\ \text{s.t.} \quad & \inf_{\tilde{\mathbf{x}}_i \sim (\mathbf{x}_i, \mathbf{I})} \mathbb{P}[y_i(\mathbf{w}^\top \tilde{\mathbf{x}}_i + b) \geq 1 - \xi_i] \geq 1 - \kappa,\end{aligned}$$

where $\tilde{\mathbf{x}}_i \sim (\mathbf{x}_i, \mathbf{I})$ denotes a family of distributions which have a common mean $\mathbf{x}_i$ and covariance $\mathbf{I}$, and $\kappa \triangleq a^2/(a^2 + C^2)$.

We next consider the case where one may like to impose additional constraints on $\mathbf{w}$. For instance, if the first feature is measured from a less reliable source, then an ideal classification rule should discount the importance of the first feature, by imposing a constraint like $|w_1| \leq 0.001$. Thus, the linear classification problem becomes

$$\begin{aligned}\min_{\mathbf{w},b} \quad & \frac{1}{m}\sum_{i=1}^m \mathbf{1}[y_i(\mathbf{w}^\top \mathbf{x}_i + b) \leq a] \\ \text{s.t.} \quad & \|\mathbf{w}\|_2 \leq C \\ & A\mathbf{w} \leq \mathbf{d}.\end{aligned}$$

Using the coherent loss to replace the objective function, and simplifying the resulting formulation, we obtain the following second order cone program

$$\begin{aligned}\min_{\hat{\mathbf{w}},\hat{b},h} \quad & \frac{1}{m}\sum_{i=1}^m [1 + ah - y_i(\hat{\mathbf{w}}^\top \mathbf{x}_i + \hat{b})]_+ \\ \text{s.t.} \quad & \|\hat{\mathbf{w}}\|_2 - Ch \leq 0 \\ & A\hat{\mathbf{w}} \leq \mathbf{d}h \\ & h > 0.\end{aligned}$$

Finally, we remark that the coherent loss approach can be kernelized, since a representation theorem (Schölkopf & Smola, 2002) still holds if the coherent loss function is used.

EXAMPLE: MULTI-CLASS SVM

The coherent loss function can also be applied in multi-class classification problems. The main idea of previous approaches (Liu & Shen, 2006; Lee et al., 2004; Crammer & Singer, 2002) of multi-class SVMs is solving one single regularization problem by imposing a penalty on the values of $f_y(\mathbf{x}) - f_z(\mathbf{x})$ for sample $(\mathbf{x}, y)$ where $f_y(\cdot)$ and $f_z(\cdot)$ are decision function for class y and z, respectively. Suppose that the training samples are drawn from k different classes and the decision function $f_y(\mathbf{x}) = \mathbf{w}_y^\top \mathbf{x} + b_y$ for each $y = 1, \cdots, k$. Consider the following 0-1 loss penalty formulation:

$$\begin{aligned}\min_{f_i} \quad & \frac{1}{m}\sum_{i=1}^m \mathbf{1}\left[\min_{z\in[k], z\neq y_i} \{f_{y_i}(\mathbf{x}_i) - f_z(\mathbf{x}_i)\} \leq a\right] \\ \text{s.t.} \quad & G_i(\mathbf{w}_i) \leq C;\ i = 1, \cdots, k \\ & \sum_{i=1}^k f_i = 0,\end{aligned}$$

where $\sum_{i=1}^k f_i = (\sum_{i=1}^k \mathbf{w}_i, \sum_{i=1}^k b_i)$, $G_i(\cdot)$ is convex (e.g. $G_i(\cdot) = \|\cdot\|_2$) and margin $a \geq 0$, then we can apply the coherent loss function approach to make an approximation:

$$\begin{aligned}\min_{f_i,\gamma>0} \quad & \frac{1}{m}\sum_{i=1}^m \left[1 - \frac{\min_{z\in[k], z\neq y_i}\{f_{y_i}(\mathbf{x}_i) - f_z(\mathbf{x}_i)\} - a}{\gamma}\right]_+ \\ \text{s.t.} \quad & G_i(\mathbf{w}_i) \leq C;\ i = 1, \cdots, k \\ & \sum_{i=1}^k f_i = 0,\end{aligned}$$

which can be simplified as the following:

$$\begin{aligned}\min_{\hat{f}_i,h>0} \quad & \frac{1}{m}\sum_{i=1}^m \left[1 + ah + \max_{z\in[k], z\neq y_i}\{\hat{f}_z(\mathbf{x}_i) - \hat{f}_{y_i}(\mathbf{x}_i)\}\right]_+ \\ \text{s.t.} \quad & hG_i(\hat{\mathbf{w}}_i/h) \leq hC;\ i = 1, \cdots, k \\ & \sum_{i=1}^k \hat{f}_i = 0,\end{aligned}$$

where $\hat{f}_i(\mathbf{x}) = \hat{\mathbf{w}}_i^\top \mathbf{x} + \hat{b}_i$. Clearly, this is a convex optimization problem and can be solved efficiently.

4. Statistical Property

In this section, we provide a statistical interpretation of minimizing the coherent loss function. As standard in learning theory, we assume that the training samples are drawn IID from an unknown distribution $\mathbb{P}$, and the goal is to find a predictor $f(\cdot)$ such that the classification error of f given below is as small as possible:

$$L(f(\cdot)) = \mathbb{E}_{(\tilde{\mathbf{x}},\tilde{y})\sim\mathbb{P}}[I(f(\tilde{\mathbf{x}}), \tilde{y})].$$

Here $(\tilde{\mathbf{x}}, \tilde{y}) \sim \mathbb{P}$ means sample $(\tilde{\mathbf{x}}, \tilde{y})$ follows the distribution $\mathbb{P}$, and $I(f(\tilde{\mathbf{x}}), \tilde{y}) = \mathbf{1}[\tilde{y}f(\tilde{\mathbf{x}}) \leq 0]$. Recall that minimizing the coherent loss function is equivalent to minimizing the following function

$$\Phi(\mathbf{u}) \triangleq \min_{\gamma>0} \frac{1}{m}\sum_{i=1}^m \phi_\gamma(u_i),$$

where $\phi_\gamma(u) = \max\{0, 1 - u/\gamma\}$. Let $\eta(\mathbf{x}) = \mathbb{P}[\tilde{y} = 1|\tilde{\mathbf{x}} = \mathbf{x}]$, then the optimal Bayes error $L^* = L(2\eta(\cdot) - 1)$.

We now develop an upper bound of the difference between $L(f(\cdot))$ and L^* using similar techniques in Zhang (2004).

For fixed γ, denote the expected loss of $f(\cdot)$ w.r.t $\phi_\gamma(\cdot)$ by

$$Q_\gamma(f(\cdot)) = \mathbb{E}_{(\tilde{\mathbf{x}},\tilde{y})\sim\mathbb{P}}[\phi_\gamma(\tilde{y}f(\tilde{\mathbf{x}}))],$$

and define

$$\begin{aligned} Q_\gamma(\eta, f) &= \eta\phi_\gamma(f) + (1-\eta)\phi_\gamma(-f) \\ \Delta Q_\gamma(\eta, f) &= Q_\gamma(\eta, f) - Q_\gamma(\eta, f^*_\gamma(\eta)), \end{aligned}$$

where $f^*_\gamma(\eta) = \arg\min_f Q_\gamma(\eta, f)$. Recall that $\phi_\gamma(u) = \max\{0, 1-u/\gamma\}$, which implies $f^*_\gamma(\eta) = \text{sign}(2\eta-1)\gamma$. Then we have the following lemma.

Lemma 1. *For $\gamma > 0$, we have $\Delta Q_\gamma(\eta, 0) = |2\eta - 1|$.*

Proof. From the definition of $Q_\gamma(\eta, f)$ and $\Delta Q_\gamma(\eta, f)$, we have

$$\begin{aligned} &\Delta Q_\gamma(\eta, f) \\ =&\eta(\phi_\gamma(f) - \phi_\gamma(f^*_\gamma(\eta))) + (1-\eta)(\phi_\gamma(-f) - \phi_\gamma(-f^*_\gamma(\eta))) \\ =&\eta\max\{0, 1-f/\gamma\} + (1-\eta)\max\{0, 1+f/\gamma\} - \\ &\eta(\max\{0, 1-\text{sign}(2\eta-1)\}) - \\ &(1-\eta)(\max\{0, 1+\text{sign}(2\eta-1)\}) \\ =&\eta\max\{0, 1-f/\gamma\} + (1-\eta)\max\{0, 1+f/\gamma\} \\ &-1+|2\eta-1|. \end{aligned}$$

This implies that $\Delta Q_\gamma(\eta, 0) = |2\eta - 1|$. □

By applying the lemma above, we can bound the classification error of $f(\cdot)$ w.r.t $\phi_\gamma(\cdot)$ in terms of $\mathbb{E}_{\tilde{\mathbf{x}}}\Delta Q_\gamma(\eta(\tilde{\mathbf{x}}), f(\tilde{\mathbf{x}}))$.

Theorem 7. *For any $\gamma > 0$ and any measurable function $f(x)$, we have*

$$\begin{aligned} L(f(\cdot)) - L^* &\le \mathbb{E}_{\tilde{\mathbf{x}}}\Delta Q_\gamma(\eta(\tilde{\mathbf{x}}), f(\tilde{\mathbf{x}})) \\ &= \mathbb{E}_{\tilde{\mathbf{x}}}[Q_\gamma(\eta(\tilde{\mathbf{x}}), f(\tilde{\mathbf{x}})) + |2\eta(\tilde{\mathbf{x}}) - 1| - 1]. \end{aligned}$$

Proof. By definition of $L(\cdot)$, it is easy to verify that

$$\begin{aligned} &L(f(\cdot)) - L(2\eta(\cdot) - 1) \\ =&\mathbb{E}_{\eta(X)\ge 0.5, f(X)<0}(2\eta(X) - 1) + \\ &\mathbb{E}_{\eta(X)<0.5, f(X)\ge 0}(1 - 2\eta(X)) \\ \le&\mathbb{E}_{(2\eta(X)-1)f(X)\le 0}|2\eta(X) - 1|. \end{aligned}$$

From Lemma 1 $\Delta Q_\gamma(\eta, 0) = |2\eta - 1|$, we have

$$L(f(\cdot)) - L^* \le \mathbb{E}_{(2\eta(\tilde{\mathbf{x}})-1)f(\tilde{\mathbf{x}})\le 0}\Delta Q_\gamma(\eta(\tilde{\mathbf{x}}), 0).$$

To complete the proof, since $\Delta Q_\gamma(\eta, f) = Q_\gamma(\eta, f) - Q_\gamma(\eta, f^*_\gamma(\eta))$, it suffices to show that $Q_\gamma(\eta(\mathbf{x}), 0) \le Q_\gamma(\eta(\mathbf{x}), f(\mathbf{x}))$ for all $\mathbf{x}$ such that $(2\eta(\mathbf{x}) - 1)f(\mathbf{x}) \le 0$. To see this, we consider three scenarios:

- $\eta > 0.5$: We have $f^*_\gamma(\eta) = \text{sign}(2\eta-1)\gamma > 0$. In addition, $(2\eta-1)f \le 0$ implies that $f \le 0$. Since $0 \in [f, f^*_\gamma(\eta)]$ and the convexity of $Q_\gamma(\eta, f)$ w.r.t. f, we have $Q_\gamma(\eta, 0) \le \max\{Q_\gamma(\eta, f), Q_\gamma(\eta, f^*_\gamma(\eta))\} = Q_\gamma(\eta, f)$.
- $\eta < 0.5$: In this case we have $f^*_\gamma(\eta) < 0$ and $f \ge 0$, which leads to $0 \in [f^*_\gamma(\eta), f]$, which implies that $Q_\gamma(\eta, 0) \le \max\{Q_\gamma(\eta, f), Q_\gamma(\eta, f^*_\gamma(\eta))\} = Q_\gamma(\eta, f)$.
- $\eta = 0.5$: Note that $f^*_\gamma = 0$, which implies that $Q_\gamma(\eta, 0) \le Q_\gamma(\eta, f)$ for all f.

From the proof of Lemma 1, we have $\Delta Q_\gamma(\eta, f) = Q_\gamma(\eta, f) + |2\eta - 1| - 1$. Hence the theorem holds. □

Corollary 1. *For any measurable function $f(x)$,*

$$\begin{aligned} L(f(\cdot)) - L^* \le \min_{\gamma>0} \mathbb{E}_{\tilde{\mathbf{x}}}[Q_\gamma(\eta(\tilde{\mathbf{x}}), f(\tilde{\mathbf{x}})) + \\ |2\eta(\tilde{\mathbf{x}}) - 1| - 1]. \end{aligned} \tag{13}$$

Proof. Since Theorem 7 holds for any $\gamma > 0$, we obtain this corollary. □

For the training samples $\{\mathbf{x}_i, y_i\}_{i=1}^m$, since $\eta(\mathbf{x}_i) = y_i \in \{1, -1\}$, the empirical estimation of the bound in (13) is $\Phi(\mathbf{u}) = \min_{\gamma>0} \frac{1}{m}\sum_{i=1}^m \phi_\gamma(u_i)$ where $u_i = y_i f(\mathbf{x}_i)$, which implies that minimizing the coherent loss function is equivalent to minimizing the empirical bound of the difference between $L(f(\cdot))$ and L^*.

5. Simulations

We report some numerical simulation results in this section to illustrate the proposed approach. Besides the regularization constraints (e.g. $\|\mathbf{w}\| \le C$ for binary-class SVMs and $\|\mathbf{w}_i\| \le C, i = 1, \cdots, k$ for multi-class SVMs), we consider the case where additional linear constraints are also imposed on the coefficient $\mathbf{w}$. For clarity, we choose a simple additional constraint $\|\mathbf{A}\mathbf{w}\|_\infty \le T$ to compare the performance of the cumulative loss formulation (SVM) and our coherent loss formulation (CCLF) for binary-class and multi-class classification, where $\mathbf{A} = [\mathbf{I}_k, \mathbf{0}] \in \mathbb{R}^{k\times n}$. In other words, the constraint ensures that the maximum of the first k elements of $\mathbf{w}$ is bounded by T. We now compare their performance under two cases: 1) k is fixed, T varies; 2) T is fixed, k varies.

Three binary-class datasets "Breast cancer", "Ionosphere" and "Diabetes", and two multi-class datasets "Wine" and "Iris" from UCI (Asuncion & Newman, 2007) are used, where we randomly pick 50% as training samples, 20% as validation samples, and the rest as testing samples. For the cumulative loss formulation approach, parameter C is

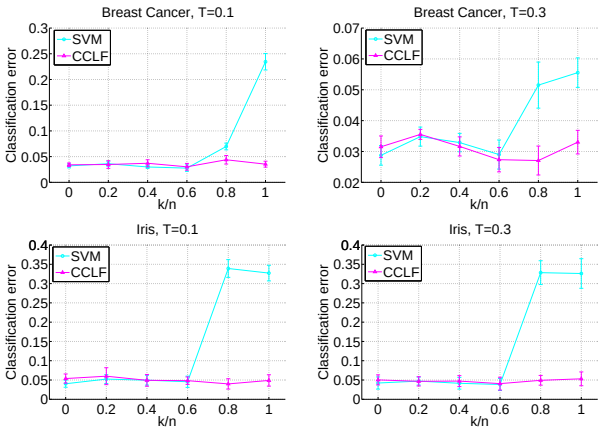

Figure 2. Performance comparison of *cumulative loss approach* vs *coherent loss approach* where bound T is fixed and the fraction k/n varies from 0.0 to 1.0. Left and right columns report the classification errors for the two cases $T = 0.1$ and $T = 0.3$.

determined by cross-validation. For the coherent loss formulation approach, parameter C is fixed while parameter a is determined by cross-validation. For each T, we repeated the experiments 20 times and computed the average classification errors. To solve the resulting optimization problems, we use CVX (Grant & Boyd, 2011; 2008), and Gurobi (Gurobi Optimization, 2013) as the solver.

Figure 3 shows the simulation results under fixed k. Clearly, when additional constraints are imposed, it appears that the coherent loss approach consistently outperforms the cumulative loss approach. When T is small, the cumulative loss approach performs much worse. When T becomes large, its performance can be close to the coherent loss approach. Figure 2 provides the results under fixed T, which shows that the coherent loss and cumulative loss approaches have similar performance when k/n is small but the coherent loss approach outperforms the cumulative loss approach when k/n is large. We believe that these phenomena are due to the fact that the coherent loss is a better approximation for the empirical classification error.

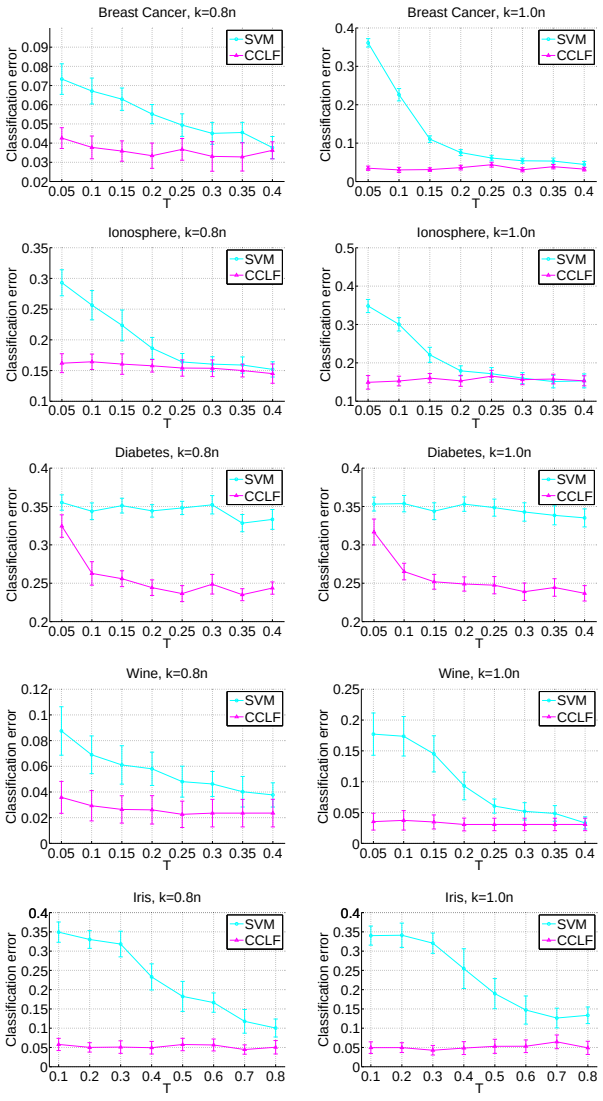

Figure 3. Performance comparison of *cumulative loss approach* vs *coherent loss approach*. Left and right columns report the classification errors for the two cases $k = 0.8n$ and $k = n$ (recall that k and n are the numbers of the rows and columns of matrix $\mathbf{A}$, respectively). The four rows, from top to bottom, report results for *Breast Cancer*, *Ionosphere*, *Diabetes*, *Wine* and *Iris*, respectively.

6. Conclusion

In this paper, we revisit the standard cumulative-loss approach in dealing with the non-convexity of the 0-1 loss function in classification, namely minimizing the sum of convex surrogates for each sample. We propose the notion of *coherent loss*, which is a tractable upper-bound of the total classification error for the entire sample set. This approach yields a strictly tighter approximation to the 0-1 loss than any cumulative loss, while preserving the tractability of the resulting optimization problem. The formulation obtained by applying the coherent loss to binary classification also has a robustness interpretation, which builds a strong connection between the coherent loss and robust SVMs. Finally, we remark that the coherent loss approach has favorable statistical properties and the simulation results show that it can outperform the standard SVM when additional constraints are imposed.

Acknowledgments

This work is partially supported by the Ministry of Education of Singapore through AcRF Tier Two grant R-265-000-443-112.

References

Arora, S., Babai, L., Stern, J., and Sweedyk, Z. The hardness of approximate optima in lattices, codes, and systems of linear equations. *Journal of Computer and System Sciences*, 54:317–331, 1997.

Asuncion, A. and Newman, D. J. UCI machine learning repository, 2007. URL http://www.ics.uci.edu/{\char126}mlearn/{MLR}epository.html.

Ben-David, S., Eiron, N., and Long, P. M. On the difficulty of approximately maximizing agreements. *Journal of Computer and System Sciences*, 66:496–513, 2003.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory*, pp. 144–152, New York, NY, 1992.

Boyd, S. and Vandenberghe, L. *Convex Optimization*. Cambridge University Press, 2004.

Brown, D.B. and Sim, M. Satisficing measures for analysis of risky positions. *Management Science*, 55(1):71–84, 2009.

Cortes, C. and Vapnik, V. N. Support vector networks. *Machine Learning*, 20:1–25, 1995.

Crammer, K. and Singer, Y. On the algorithmic implementation of multiclass kernel-based vector machines. *J. Mach. Learn. Res.*, 2:265–292, 2002.

Freund, Y. and Schapire, R. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1): 119–139, 1997.

Friedman, J., Hastie, T., and Tibshirani, R. Additive logistic regression: A statistical view of boosting. *Annals of Statistics*, 28:337–407, 2000.

Grant, M. and Boyd, S. Graph implementations for nonsmooth convex programs. In *Recent Advances in Learning and Control*, pp. 95–110. Springer-Verlag Limited, 2008.

Grant, M. and Boyd, S. CVX: Matlab software for disciplined convex programming, version 1.21. http://cvxr.com/cvx, 2011.

Gurobi Optimization, Inc. Gurobi optimizer reference manual, 2013. URL http://www.gurobi.com.

Lee, Y., Lin, Y., and Wahba, G. Multicategory support vector machines, theory, and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association*, 99:67–81, 2004.

Liu, Y. and Shen, X. Multicategory φ-learning. *Journal of the American Statistical Association*, 101(474):500–509, 2006.

Poggio, T., Rifkin, R., Mukherjee, S., and Niyogi, P. General conditions for predictivity in learning theory. *Nature*, 428(6981):419–422, 2004.

Schapire, E. and Singer, Y. Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 37:297–336, 1999.

Schölkopf, B. and Smola, A. J. *Learning with Kernels*. MIT Press, 2002.

Shivaswamy, P. K., Bhattacharyya, C., and Smola, A. J. Second order cone programming approaches for handling missing and uncertain data. *Journal of Machine Learning Research*, 7:1283–1314, July 2006.

Vapnik, V. N. and Chervonenkis, A. The necessary and sufficient conditions for consistency in the empirical risk minimization method. *Pattern Recognition and Image Analysis*, 1(3):260–284, 1991.

Vapnik, V. N. and Lerner, A. Pattern recognition using generalized portrait method. *Automation and Remote Control*, 24:744–780, 1963.

Zhang, T. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32:56–85, 2004.