Adaptive Confidence Intervals for the Test Error in Classification

Eric B. Laber and Susan A. Murphy[1]

Abstract

The estimated test error of a learned classifier is the most commonly reported measure of classifier performance. However, constructing a high quality point estimator of the test error has proved to be very difficult. Furthermore, common interval estimators (e.g. confidence intervals) are based on the point estimator of the test error and thus inherit all the difficulties associated with the point estimation problem. As a result, these confidence intervals do not reliably deliver nominal coverage. In contrast we directly construct the confidence interval by use of smooth data-dependent upper and lower bounds on the test error. We prove that for linear classifiers, the proposed confidence interval automatically adapts to the non-smoothness of the test error, is consistent under fixed and local alternatives, and does not require that the Bayes classifier be linear. Moreover, the method provides nominal coverage on a suite of test problems using a range of classification algorithms and sample sizes.

Keywords: Classification, Test Error, Pretesting, Confidence Intervals, Non-Regular Asymptotics

1 Introduction

In classification problems, we observe a training set of (feature, label) pairs, $\mathcal{T} = \{(X_i, Y_i)\}_{i=1}^n$. The goal is use this sample to construct a classifier, say $\hat{c}$, so that when presented with a

[1] Eric Laber is in the Department of Statistics at the University of Michigan, Ann Arbor, MI, 48109 (E-mail: *laber@umich.edu*). Susan A. Murphy is Professor of Statistics and Psychiatry, University of Michigan, Ann Arbor, MI, 48109 (E-mail: *samurphy@umich.edu*). This research is supported by NIH grants R01 MH080015 and P50 DA10075. The authors thank Min Qian, Zhigou Li, Diane Lambert, Kerby Shedden, and Vijay Nair for insightful comments and criticisms. In addition, the authors wish to thank the Editor and anonymous reviewers for criticisms and insights that made for a much better paper.

new feature, X, $\hat{c}(X)$ will accurately predict the unobserved label, Y. Accurate prediction corresponds to small test error; recall that the test error is given by $\tau(\hat{c}) = P1_{\hat{c}(X)\neq Y}$ where $P1_{\hat{c}(X)\neq Y} = \int 1_{\hat{c}(x)\neq y} dP(x, y)$ denotes expectation over the distribution P of (X, Y) only, and not the distribution of the training set. The test error $\tau(\hat{c})$ is a functional of $\hat{c}$ and thus is a random quantity. For this reason $\tau(\hat{c})$ is sometimes referred to as the conditional test error (Efron 1997; Hastie et al. 2009; Chung and Han 2009). Estimation of the test error typically employs resampling. Most commonly, the leave-one-out or k-fold cross-validated test error is reported in practice. Bootstrap estimates of the test error were suggested by Efron (1983) and later refinements were given by Efron and Tibshirani (1995, 1997). There have been a number of simulation studies comparing these approaches; some references include (Efron 1983; Chernick et al. 1985; Kohavi 1995; Krzanowksi and Hand 1996). A nice survey of estimators is given by Schiavo and Hand (2000). However many have documented that estimators of the test error are plagued by bias and high variance across training sets (Zhang 1995; Isaakson 2008; Hastie et al. 2009) and consequently the test error is accepted to be a difficult quantity to estimate accurately. Two reasons for this problematic behavior are that some classification algorithms result in a $\hat{c}$ that is a non-smooth functional of the training set, and, even when $\hat{c}$ is a smooth functional of the training set, the test error is the expectation of a non-smooth function of $\hat{c}$.

An alternative to point estimation is interval estimation (e.g. a confidence interval). However, this approach has also been problematic likely because researchers have followed what we call the "point estimation paradigm": as a first step a point estimator of the test error is constructed, and as a second step, the distribution of this estimator is approximated. The problem with this approach is that a problematic point estimator of the test error makes the second step very difficult. The point estimation paradigm was employed by Efron and Tibshirani (1997) where the standard error of their smoothed leave-one-out estimator was approximated using the nonparametric delta method. Efron and Tibshirani noted that this approach would not work, however, for their more refined .632 (or .632+) estimators because of non-smoothness. Yang (2006) follows this paradigm as well, using a normal approximation to the repeated split cross-validation estimator. In practice, the point estimation paradigm is often applied by simply bootstrapping the estimator of the test error (see Jiang et al., 2008; Chung and Han 2009). These methods, while intuitive, lack theoretical justification.

We consider interval estimators for linear classifiers constructed from training sets in which the number of features is less than the training set size ($p << n$). As will be seen, even in this simple setting, natural approaches to constructing interval estimators for the test error can perform poorly. Instead of using the point estimation paradigm, we directly construct the confidence interval by use of smooth data-dependent upper and lower bounds on the test error. These bounds are sufficiently smooth so that their bootstrap distribution can be used to construct valid confidence intervals. Moreover, these bounds are adaptive in the sense that under certain settings exact coverage is delivered.

The outline of this paper is as follows. In Section 2 we illustrate the small sample problems that motivate the use of approximations in a non-regular asymptotic framework. Section 3 introduces the Adaptive Confidence Interval (ACI). The ACI is shown to be consistent under fixed and local alternatives. Section 4 addresses the computational issues involved in constructing the ACI. A computationally efficient (polynomial time) convex relaxation of the ACI is developed and shown to provide nearly identical results to exact computation. Section 5 provides a large experimental study of the ACI and several competitors. A variety of classifiers and sample sizes are considered on a suite of ten examples. The ACI is shown to provide correct coverage while being shorter in length than competing methods. Section 6 discusses a number of generalizations and directions for future research. Most proofs are left to the online supplement.

2 Motivation

Throughout we assume that the training set is an *iid* sample $\mathcal{T} = \{(X_i, Y_i)\}_{i=1}^n$ drawn from some unknown joint distribution P. The features X are assumed to take values in $\mathbb{R}^p$ while the labels are coded $Y \in \{-1, 1\}$. To construct the linear classifier we fit a linear model $\hat{f}_{\mathcal{T}}(x) = x^t\hat{\beta}_n$ by minimizing a convex criterion function. That is, we construct $\hat{\beta}_n \triangleq \arg\min_{\beta\in\mathbb{R}^p} \mathbb{P}_n L(X, Y, \beta)$ where $\mathbb{P}_n$ is the empirical measure and $L(X, Y, \beta)$ is a convex function of β (e.g., hinge loss with an L_2 penalty in the case of linear support vector machines). The classifier is the sign of the linear fit; that is, the predicted label y at input x is assigned according to $\hat{c}(x) = sign(x^t\hat{\beta}_n)$ (define $sign(0) = 1$). Recall that the test error of the learned classifier is defined as

$$\tau(\hat{c}) \triangleq P1_{sign(X^t\hat{\beta}_n)\neq Y} = P1_{YX^t\hat{\beta}_n<0},$$

where P denotes expectation with respect to X and Y.

As discussed in the introduction, the test error is a non-smooth functional of the training data. To see this and to gain a clearer understanding of the test error note

$$\tau(\hat{c}) = const + \int_{\mathbb{R}^p} [2q(x) - 1]\, 1_{x^t\hat{\beta}_n<0} dP_X(x), \tag{1}$$

where $q(x) \triangleq P(Y = 1|X = x)$. Recall that $sign(2q(x) - 1)$ is the Bayes classifier. Then

$$Var\left(\tau(\hat{c})\right) = \mathbb{E}\left(\int_{\mathbb{R}^p} [2q(x) - 1]\left(1_{x^t\hat{\beta}_n<0} - \mathbb{E}1_{x^t\hat{\beta}_n<0}\right) dP_X(x)\right)^2, \tag{2}$$

where $\mathbb{E}$ denotes the expectation over *iid* training sets of size n drawn from P. The form of $Var\left(\tau(\hat{c})\right)$ reveals that there are two scenarios in which $\tau(\hat{c})$ is highly variable. The first occurs when $x^t\hat{\beta}_n$ is likely to be small relative to $Var(x^t\hat{\beta}_n)$ over a *large* range of x where

$q(x) \neq 1/2$. Notice that this might occur when the classifier does well but is subject to overfitting. The second scenario occurs when $x^t\hat{\beta}_n$ is likely to be small relative to $Var(x^t\hat{\beta}_n)$ over a *small* range of x where $q(x)$ is far from $1/2$. In this scenario there may be little overfitting but the classifier may be far from the Bayes rule and hence of poor quality. Note that poor classifier performance and overfitting are hallmarks of small samples. In either case, $\tau(\hat{c})$ need not concentrate around $\mathbb{E}\tau(\hat{c})$.

In order to provide good intuition for the small sample case, we require an asymptotic framework wherein the test error $\tau(\hat{c})$ does not concentrate about $\mathbb{E}\tau(\hat{c})$, even in large samples. One way of achieving this is to permit $P(X^t\beta^* = 0)$ to be positive where $\beta^* \triangleq \arg\min_{\beta\in\mathbb{R}^p} PL(X, Y, \beta)$. This ensures that for all $x \in \mathbb{R}^p$ that satisfy $x^t\beta^* = 0$, the indicator function $1_{x^t\hat{\beta}_n<0} = 1_{x^t\sqrt{n}(\hat{\beta}_n-\beta^*)<0}$ never settles down to a constant but rather converges to a non-degenerate distribution. Furthermore, if for a non-null subset of these x's we have $q(x) \neq 1/2$, then $Var\left(\tau(\hat{c})\right)$ does not converge to zero. Hereafter we refer to this as the *non-regular framework*. This language is consistent with that of Bickel et al. (2001). However, unlike the usual notion of non-regularity the limiting distribution of $\sqrt{n}(\hat{\tau}(\hat{c}) - \tau(\hat{c}))$ depends not only on the value of β^* but also the marginal distribution of X.

To see why it is useful to consider approximations that are valid even in the non-regular asymptotic framework we consider simulated data, which we call the quadratic example. Here the generative model satisfies $P(X^t\beta^* = 0) = 0$. Data are generated according to the following mechanism

$$\begin{aligned} X_1, X_2 &\sim_{iid} Unif[0,5] \\ \epsilon &\sim N(0, 1/4) \\ Y &= sign(X_2 - (4/25)X_1^2 - 1 + \epsilon). \end{aligned}$$

The working classifier is given by $\hat{c}(x) = sign(\hat{\beta}_{n0} + \hat{\beta}_{n1}x_1 + \hat{\beta}_{n2}x_2)$ where $\hat{\beta}_n$ is constructed using squared error loss $L(X, Y, \beta) \triangleq (1 - YX^t\beta)^2$. In this example $\beta^* \approx (-.225, -317, .439)$ so that the continuity of X_1 and X_2 ensures that the regularity condition $P(X^t\beta^* = 0) = 0$ is satisfied. Consider two seemingly reasonable, and commonly employed methods for constructing a confidence set. The first is the centered percentile bootstrap (CPB). The CPB confidence set is formed by bootstrapping the centered and scaled in-sample error $\sqrt{n}(\mathbb{P}_n - P)1_{YX^t\hat{\beta}_n<0}$. Note that $\sqrt{n}(\mathbb{P}_n - P)1_{YX^t\hat{\beta}_n<0} = \sqrt{n}(\hat{\tau}(\hat{c}) - \tau(\hat{c}))$ where $\hat{\tau}(\hat{c}) \triangleq \mathbb{P}_n 1_{YX^t\hat{\beta}_n<0}$ is the in-sample error. More specifically, let $\hat{u}$ and $\hat{l}$ be the $1 - \gamma/2$ and $\gamma/2$ percentiles of

$$\sqrt{n}(\hat{\mathbb{P}}_n^{(b)} - \mathbb{P}_n)1_{YX^t\hat{\beta}_n^{(b)}<0}, \tag{3}$$

where $\hat{\mathbb{P}}_n^{(b)} \triangleq n^{-1}\sum_{i=1}^n M_{ni}\delta_{(x_i,y_i)}$ is the bootstrap empirical measure with weights $(M_{n1}, M_{n2}, \ldots, M_{nn}) \sim Multinomial(n, \frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n})$ and $\hat{\beta}_n^{(b)} \triangleq \arg\min_{\beta\in\mathbb{R}^p} \hat{\mathbb{P}}_n^{(b)} L(X, Y, \beta)$. Then the $1 - \gamma$ CPB interval is given by $[\hat{\tau}(\hat{c}) - \hat{u}/\sqrt{n}, \hat{\tau}(\hat{c}) - \hat{l}/\sqrt{n}]$. The second approach is based on the asymptotic approximation

$$\sqrt{n}(\mathbb{P}_n - P)1_{YX^t\hat{\beta}_n<0} \approx N\left(0, (1 - P1_{YX^t\beta^*<0})P1_{YX\beta^*<0}\right). \tag{4}$$

Thus the normal approximation confidence set is given by $\hat{\tau}(\hat{c}) \pm z_{1-\gamma}\sqrt{\frac{\hat{\tau}(\hat{c})(1-\hat{\tau}(\hat{c}))}{n}}$ (see the binomial approximation in Chung and Han 2009). If $P(X^t\beta^* = 0) = 0$ then both methods can be shown to be consistent.

The left hand side of Figure 1 shows the estimated coverage using 1000 Monte Carlo iterations of the CPB with 1000 bootstrap resamples, and the normal approximation. Both methods severely undercover in small samples. This is especially troubling since (i) the problem is low-dimensional, (ii) the linear classifier is of relatively high quality, (for example if $n = 30$ the expected test error $\mathbb{E}\tau(\hat{c}) \approx .11$) and (iii) the regularity condition $P(X^t\beta^* = 0) = 0$ is satisfied. Why do these methods fail? Neither method correctly captures the additional variation in the test error across training samples due to the non-smoothness of the test error. Since the generative model satisfies the condition $P(X^t\beta^* = 0) = 0$, the variation across training sets eventually becomes negligible and the methods deliver the desired coverage for n large.

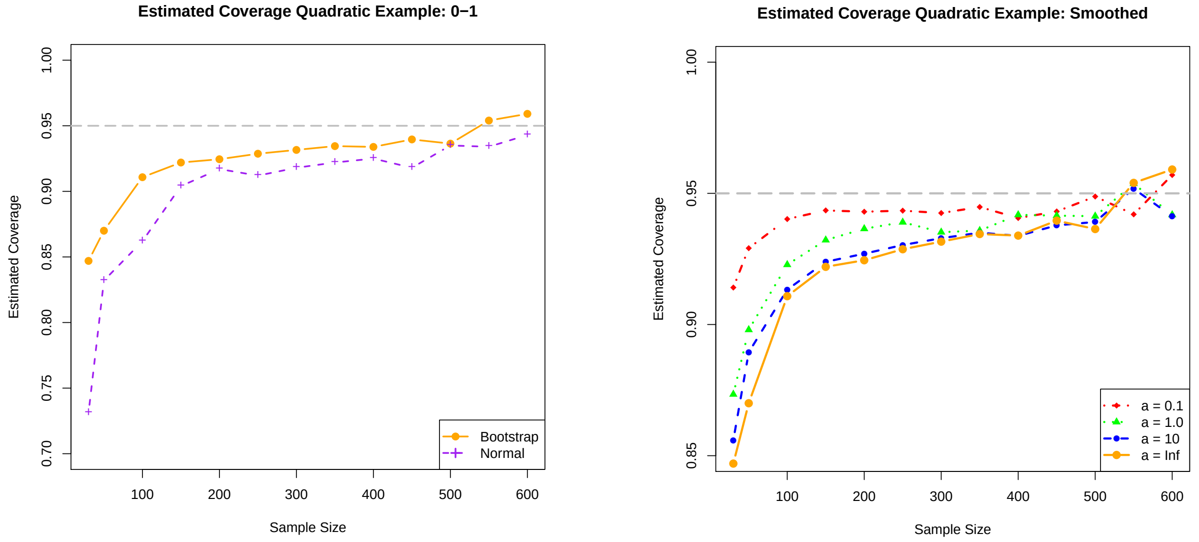

Figure 1: Left: Coverage of centered percentile bootstrap and normal approximations for constructing confidence sets for $\tau(\hat{c})$. Right: Coverage of centered percentile bootstrap with smoothed target $\tau_{smoothed}(\hat{c}) \triangleq P(1 + exp(aY\hat{c}(X)))^{-1}$ for varying values of a; a value of $a = \infty$ corresponds to $\tau(\hat{c})$. Results are based on 1000 Monte Carlo iterations, target coverage is .950. The performance of the ACI on this example can found in Section 5 under the example labeled "quad."

To illustrate the effect of non-smoothness on the coverage consider the problem of finding a confidence interval for the functional $\tau_{smoothed}(\hat{c}) \triangleq P(1 + exp(aY\hat{c}(X)))^{-1}$, where a is a positive free parameter. Notice that the size of a varies inversely with the smoothness of $\tau_{smooth}(\hat{c})$. A value of $a > 0$ gives the expectation of a sigmoid function and a value of $a = \infty$ corresponds to $\tau(\hat{c})$. Coverage for $a = 0.1, 1.0$, and 10 are given in the right hand side of Figure 1. Notice that coverage increases with the smoothness of the target $\tau_{smoothed}(\hat{c})$. The dramatic difference in coverage between $a = .1$ and $a = \infty$ suggests that a large component of the anti-conservatism is indeed attributable to non-smoothness.

Operating in the regular framework there is no indication that these methods may not work well. In the non-regular framework, however, both of these methods are inconsistent. To see this in the case of the CPB, write

$$\begin{aligned}\sqrt{n}(\hat{\mathbb{P}}_n^{(b)} - \mathbb{P}_n)1_{YX^t\hat{\beta}_n^{(b)}<0} &= \sqrt{n}(\hat{\mathbb{P}}_n^{(b)} - \mathbb{P}_n)1_{X^t\beta^*=0}1_{YX^t\left[\sqrt{n}(\hat{\beta}_n^{(b)}-\hat{\beta}_n)+\sqrt{n}(\hat{\beta}_n-\beta^*)\right]<0} \\ &\quad + \sqrt{n}(\hat{\mathbb{P}}_n^{(b)} - \mathbb{P}_n)1_{X^t\beta^*\neq 0}1_{YX^t\hat{\beta}_n^{(b)}<0}. \qquad (5)\end{aligned}$$

The first term on the right hand side of (5) appears because we allow $P(X^t\beta^* = 0) > 0$ in the non-regular framework; conditioned on the data the term $\sqrt{n}(\hat{\beta}_n - \beta^*)$ does not have a limit and consequently the CPB is inconsistent. A detailed proof is omitted (see for example Shao 1994). The inconsistency of the normal approximation can be seen by examining the limiting distribution of $\sqrt{n}(\mathbb{P}_n - P)1_{YX^t\hat{\beta}_n<0}$ in the non-regular framework. This limit is given in Theorem 3.1.

3 Adaptive confidence interval

In this section we introduce our method for constructing a confidence interval for the test error. This section is organized as follows. We begin by constructing adaptive confidence interval. Next, we establish the theoretical underpinnings of the method under fixed alternatives. Following this we provide a (heuristic) justification for our method using local alternatives. Finally, we discuss the choice of a tuning parameter required by the method.

3.1 Construction of the ACI

We propose an method of constructing a confidence interval that is consistent in the non-regular framework. We refer to this method as the Adaptive Confidence Interval (ACI) because, it is adaptive in two ways. First, unlike the CPB, the ACI provides asymptotically valid confidence intervals regardless of the true parameter values; intuitively the ACI achieves this by adapting to the amount of non-smoothness in the test error. Second, in settings (see Corollary 3.4) in which the CPB is consistent, the upper and lower limits of the ACI are adaptive in that these limits have the same distribution as the upper and lower limits of the CPB.

The ACI is based on bootstrapping an upper bound of the functional $\sqrt{n}(\mathbb{P}_n - P)1_{YX^t\hat{\beta}_n<0}$. This upper bound is constructed by first partitioning the training data $\mathcal{T}$ into two groups (i) points that are far from the boundary $x^t\beta^* = 0$, and (ii) points that are too close to delineate from being on the boundary. The upper bound is constructed by taking the supremum over all possible classifications of the points that we cannot distinguish from lying on the boundary. More precisely, under the non-regular framework the scaled and centered test error can be decomposed as

$$\mathbb{G}_n 1_{YX^t\hat{\beta}_n<0} = \mathbb{G}_n 1_{X^t\beta^*=0} 1_{YX^t\hat{\beta}_n<0} + \mathbb{G}_n 1_{X^t\beta^*\neq 0} 1_{YX^t\hat{\beta}_n<0}, \tag{6}$$

where $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P)$. The first term on the right hand side of (6) corresponds to points on the decision boundary $x^t\beta^* = 0$, and the second term corresponds to points that are not on this boundary. That is, the domain of X is partitioned into two-sets. We operationalize this partitioning using a series of hypothesis tests. For each $X = x$ we test $H_0 : x^t\beta^* = 0$ against a two-sided alternative. Let Σ denote the asymptotic covariance of $\hat{\beta}_n$ (see below). Then the test rejects when the statistic $\frac{(x^t\hat{\beta}_n)^2}{x^t\Sigma x}$ is large. The bounds are obtained by computing the supremum (infemum) over all classifications of points for which the test fails to reject. In particular, an upper bound on $\mathbb{G}_n 1_{YX^t\hat{\beta}_n<0}$ is given by

$$u(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n) = \sup_{b\in\mathbb{R}^p} \mathbb{G}_n 1_{\frac{(X^t\hat{\beta}_n)^2}{X^t\Sigma X} \leq \frac{1}{a_n}} 1_{YX^tb<0} + \mathbb{G}_n 1_{\frac{(X^t\hat{\beta}_n)^2}{X^t\Sigma X} > \frac{1}{a_n}} 1_{YX^t\hat{\beta}_n<0}, \tag{7}$$

and an lower bound is given by

$$\ell(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n) = \inf_{b \in \mathbb{R}^p} \mathbb{G}_n 1_{\frac{(X^t\hat{\beta}_n)^2}{X^t\Sigma X} \leq \frac{1}{a_n}} 1_{YX^tb<0} + \mathbb{G}_n 1_{\frac{(X^t\hat{\beta}_n)^2}{X^t\Sigma X} > \frac{1}{a_n}} 1_{YX^t\hat{\beta}_n<0}. \tag{8}$$

The choice of a_n, is discussed at the end of this Section. Put $b = \hat{\beta}_n$ to see that (7) and (8) are upper and lower bounds, respectively.

Suppose we want to construct a $1-\delta\%$ confidence interval for the test error. We have that

$$\mathbb{P}_n 1_{YX^t\hat{\beta}_n<0} - (1/\sqrt{n})u(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n) \leq P1_{YX^t\hat{\beta}_n<0} \leq \mathbb{P}_n 1_{YX^t\hat{\beta}_n<0} - (1/\sqrt{n})\ell(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n).$$

We approximate the distribution of $u(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n)$, $\ell(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n)$ by bootstrap. The bootstrap is shown to be consistent later in this section. Denote the $1-\delta/2$ percentile of the bootstrap distribution of $u(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n)$ by $u_{1-\delta/2}$ and the $\delta/2$ percentile of the bootstrap distribution of $\ell(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n)$ by $\ell_{\delta/2}$. The $1-\delta\%$ ACI is given by

$$\mathbb{P}_n 1_{YX^t\hat{\beta}_n<0} - (1/\sqrt{n})u_{1-\delta/2} \leq P1_{YX^t\hat{\beta}_n<0} \leq \mathbb{P}_n 1_{YX^t\hat{\beta}_n<0} - (1/\sqrt{n})\ell_{\delta/2}. \tag{9}$$

3.2 Properties of the ACI

In the remainder of the paper we verify that the ACI is asymptotically of the correct size even if the problem is non-regular (e.g. $P(X^t\beta^* = 0) > 0$) and we evaluate the performance of the ACI in small samples. A method for efficiently approximating the ACI is given and shown to be almost identical to exact computation on a suite of examples. Most proofs are deferred to the online supplement.

First we provide the asymptotic distribution of $u(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n)$ and $\ell(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n)$. Throughout we make the following assumptions.

(A1) $L(X, Y, \beta)$ is convex with respect to β for each fixed $(x, y) \in \mathbb{R}^p \times \{-1, 1\}$.

(A2) $Q(\beta) \triangleq PL(X, Y, \beta)$ exists and is finite for all $\beta \in \mathbb{R}^p$.

(A3) $\beta^* \triangleq \arg\min_{\beta \in \mathbb{R}^p} Q(\beta)$ exists and is unique.

(A4) Let $g(X, Y, \beta)$ be a sub-gradient of $L(X, Y, \beta)$. Then $P||g(X, Y, \beta)||^2 < \infty$ for all β in a neighborhood of β^*.

(A5) $Q(\beta)$ is twice continuously differentiable at β^* and $H = \nabla^2 Q(\beta^*)$ is positive definite.

(A6) $\lim_{n\to\infty} a_n = \infty$ but $a_n = o(n)$.

These assumptions are quite mild and hold for most commonly used loss functions (e.g., exponential loss, squared error loss, hinge loss–if P has a smooth density at 1, logistic loss, etc.). Recall that a subgradient satisfies $L(x, y, \gamma) + (\beta - \gamma)^t g(x, y, \gamma) \leq L(x, y, \beta)$ for all $(x, y) \in \mathbb{R}^p \times \{-1, 1\}$ and $\gamma, \beta \in \mathbb{R}^p$. All convex functions have a measurable subgradient. Let Ω be the covariance matrix of the sub-gradient of $L(x, y, \beta)$ at β^*. Under (A1)-(A5) Haberman (1989; see also Niemiro, 1992) proved that $\hat{\beta}_n$ converges with probability one to β^* and $\sqrt{n}(\hat{\beta}_n - \beta^*)$ converges in distribution to $z_\infty =_{\mathcal{L}} N(0, H^{-1}\Omega H^{-1})$.

Let $\mathbb{V}$ be a Brownian-Bridge indexed by $\mathbb{R}^p$ with the variance-covariance function

$$Cov(\mathbb{V}(\phi), \mathbb{V}(\gamma)) = P\left[1_{X^t\beta^*=0}1_{YX^t\phi<0} - P1_{X^t\beta^*=0}1_{YX^t\phi<0}\right] \\ \times \left[1_{X^t\beta^*=0}1_{YX^t\gamma<0} - P1_{X^t\beta^*=0}1_{YX^t\gamma<0}\right]. \quad (10)$$

Furthermore, let $\mathbb{B}(\beta^*)$ denote a mean zero normal random variable with variance $P(1_{X^t\beta^*\neq 0}1_{YX^t\beta^*<0} - P1_{X^t\beta^*\neq 0}1_{YX^t\beta^*<0})^2$.

Theorem 3.1. *Let $\mathbb{V}$, $\mathbb{B}(\beta^*)$ and z_∞ be as above. Assume (A1)-(A6). Then*

1. $\mathbb{G}_n 1_{YX^t\hat{\beta}_n<0} \rightsquigarrow \mathbb{V}(z_\infty) + \mathbb{B}(\beta^*)$,

2. $u(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n) \rightsquigarrow \sup_{u\in\mathbb{R}^p} \mathbb{V}(u) + \mathbb{B}(\beta^*)$ *and* $\ell(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n) \rightsquigarrow \inf_{u\in\mathbb{R}^p} \mathbb{V}(u) + \mathbb{B}(\beta^*)$.

Note that the limiting distributions of $u(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n)$, $\ell(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n)$ and $\mathbb{G}_n 1_{YX^t\hat{\beta}_n<0}$ have the same regular component $\mathbb{B}(\beta^*)$; the three limits differ only in the non-regular component. Note also that the form of the covariance function of $\mathbb{V}$ given in (10) and the form of the limiting distribution of $u(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n)$ (or $\ell(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n)$) shows that if the margin condition $P(X^t\beta^* = 0) = 0$ holds, then $u(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n) \rightsquigarrow \mathbb{B}(\beta^*) =_{\mathcal{L}} lim_{n\to\infty}\mathbb{G}_n 1_{YX^t\hat{\beta}_n<0}$ and similarly for $\ell(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n)$. That is, if the margin condition holds, the limiting distribution of the functional used to construct the ACI is the same as the limiting distribution of the functional $\mathbb{G}_n 1_{YX^t\hat{\beta}_n<0}$. From a practical point of view this means that for problems where the regular framework is applicable, for example, if the sample size is large or points are well separated from the boundary, the ACI is asymptotically exact.

Another scenario in which the limiting distribution of $u(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n)$, $\ell(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n)$ and $\mathbb{G}_n 1_{YX^t\hat{\beta}_n<0}$ are the same is when the Bayes decision boundary is linear. In this case $q(x) = 1/2$ if $x^t\beta^* = 0$ where $q(x) = P(Y = 1|X = x)$. (Here, we assume that the loss function is classification-calibrated (Bartlett 2005). All loss functions mentioned in this paper are classification-calibrated.) Then for any fixed $u \in \mathbb{R}^p$ we have

$$\begin{aligned} P1_{X^t\beta^*=0}1_{YX^tu<0} &= \int_{\{x\,:\,x^t\beta^*=0\}} \left[q(x)1_{x^tu<0} + (1-q(x))(1-1_{x^tu<0})\right] dP_X(x) \\ &= \int_{\{x\,:\,x^t\beta^*=0\}} \left[2q(x)-1\right]1_{x^tu<0}\, dP_X(x) + \frac{1}{2}P(X^t\beta^* = 0) \\ &= \frac{1}{2}P(X^t\beta^* = 0). \end{aligned}$$

The form of the variance of $\mathbb{V}$ and the above series of equalities show that if the Bayes decision boundary is linear then $\mathbb{V}(u) =_{\mathcal{L}} N(0, \frac{1}{2}(1 - \frac{1}{2}P1_{X^t\beta^*=0})P1_{X^t\beta^*=0})$ for all $u \in \mathbb{R}^p$.

Therefore, if the Bayes decision is linear

$$\begin{aligned}
lim_{n\rightarrow\infty} u(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n) &=_{\mathcal{L}} \sup_{u\in\mathbb{R}^p} \mathbb{V}(u) + \mathbb{B}(\beta^*) \\
&=_{\mathcal{L}} N\left(0, (1 - P1_{X^t\beta^*<0})P1_{X^t\beta^*<0}\right) + \mathbb{B}(\beta^*) \\
&=_{\mathcal{L}} \mathbb{V}(z_\infty) + \mathbb{B}(\beta^*) \\
&=_{\mathcal{L}} \lim_{n\rightarrow\infty} \sqrt{n}(\mathbb{P}_n - P)1_{YX^t\hat{\beta}_n<0},
\end{aligned}$$

where the first and last equalities follow from Theorem 3.1, and the second and third equalities follow since $\mathbb{V}$ is constant across all indices. We have proved the following result.

Corollary 3.2. *Assuming (A1)-(A6) hold then if either (i) the Bayes decision boundary is* $sign(X^t\beta^*)$ *or (ii)* $P(X^t\beta^* = 0) = 0$ *then* $u(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n)$, $\ell(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n)$ *and* $\mathbb{G}_n 1_{YX^t\hat{\beta}_n<0}$ *have the same limiting distribution.*

The implication of the above theorem and corollary is that when either of the above conditions hold the ACI should provide the nominal coverage. When neither event holds then the ACI may be conservative. In simulations we shall see that the degree of conservatism is small.

The ACI in (9) utilizes a bootstrap approximation to the distribution of $u(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n)$, $\ell(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n)$. The next theorem concerns the consistency of the bootstrap distributions. Let $\hat{\Sigma}_n$ be a weakly consistent estimator of Σ (e.g. the plug-in estimator). Define $BL_1(\mathbb{R}^2)$ to be the space of bounded Lipschitz-1 functions on $\mathbb{R}^2$ and let $\mathbb{E}_M$ denote the expectation with respect to the bootstrap weights.

Theorem 3.3. *Assume (A1)-(A6). Then* $\{u(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n),\ \ell(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n)\}$ *and* $\{u(\mathbb{G}_n^{(b)}, \hat{\beta}_n^{(b)}, \hat{\Sigma}_n, a_n),\ \ell(\mathbb{G}_n^{(b)}, \hat{\beta}_n^{(b)}, \hat{\Sigma}_n, a_n)\}$ *converge to the same limiting distribution in prob-*

ability. That is,

$$\sup_{h \in BL_1(\mathbb{R}^2)} \Big| \mathbb{E} h\left(\{u(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n),\ \ell(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n)\}\right) \\ - \mathbb{E}_M h\left(\{u(\mathbb{G}_n^{(b)}, \hat{\beta}_n^{(b)}, \hat{\Sigma}_n, a_n),\ \ell(\mathbb{G}_n^{(b)}, \hat{\beta}_n^{(b)}, \hat{\Sigma}_n, a_n)\}\right) \Big|$$

converges in probability to zero.

Thus the ACI provides asymptotically valid confidence intervals. Moreover we have the following.

Corollary 3.4. *Assuming (A1)-(A6) hold then if either (i) the Bayes decision boundary is sign($X^t\beta^*$) or (ii) $P(X^t\beta^* = 0) = 0$ then $u(\mathbb{G}_n^{(b)}, \hat{\beta}_n^{(b)}, \hat{\Sigma}_n, a_n)$, $\ell(\mathbb{G}_n^{(b)}, \hat{\beta}_n^{(b)}, \hat{\Sigma}_n, a_n)$ and $\mathbb{G}_n 1_{YX^t\hat{\beta}_n < 0}$ converge to the same limiting distribution, in probability.*

Thus, the ACI is also adaptive in the sense that in settings where the centered percentile bootstrap *would* be consistent, $u(\mathbb{G}_n^{(b)}, \hat{\beta}_n^{(b)}, \hat{\Sigma}, a_n)$, $\ell(\mathbb{G}_n^{(b)}, \hat{\beta}_n^{(b)}, \hat{\Sigma}, a_n)$ and $\mathbb{G}_n 1_{YX^t\hat{\beta}_n < 0}$ have the same limiting distribution.

3.3 Local Alternatives

In Section 2 we motivated the use of a non-regular asymptotic framework in order to gain intuition for small samples. An alternative strategy for developing intuition for non-regular problems is to study the limiting behavior of $\sqrt{n}(\hat{\beta}_n - \beta^*)$ under local alternatives. This strategy has roots in Econometrics.

In econometrics, a common strategy to constructing procedures with good small sample properties in non-regular settings is to utilize alternatives local to the parameter values that cause the non-regularity (Andrews 2000; Cheng 2008; Xie 2009). To see this recall that in small samples a non-negligible proportion of the inputs x are in a $\sqrt{n}$-neighborhood of the decision boundary $x^t\beta^* = 0$ which causes the indicator function $1_{x^t\hat{\beta}_n < 0}$ to become unstable.

In the prior sections we assumed that there was a non-null probability that an input lies *exactly* on the boundary in order to retain the instability of the indicator function even in large samples. Another way to maintain this instability is by considering local alternatives.

The ACI can be seen as arising as an asymptotic approximation under local alternatives in the following way. In particular, suppose that a training set $\mathcal{T}_n = \{(X_{ni}, Y_{ni})\}_{i=1}^n$ is drawn *iid* from distribution P_n for which

$$\beta_n^* \triangleq \arg\min_{\beta\in\mathbb{R}^p} P_n L(X, Y, \beta) = \beta^* + \Gamma/\sqrt{n} \tag{11}$$

for some $\Gamma \in \mathbb{R}^p - \{0\}$. In addition, we assume that $P(X^t\beta^* = 0) > 0$ (while $P_n(X^t\beta_n^* = 0) > 0$ may or may not hold). A general tactic is to derive the limiting distribution of an estimator which will depend on the local parameter Γ and then take a supremum over this parameter to construct a confidence interval. As a first step in following this approach we might expect that

$$\begin{aligned}\mathbb{G}_n 1_{YX^t\hat{\beta}_n<0} = \mathbb{G}_n 1_{X^t\beta^*=0} 1_{YX^t\left[\sqrt{n}(\hat{\beta}_n-\beta_n^*)+\Gamma\right]<0} + \mathbb{G}_n 1_{X^t\beta^*\neq 0} 1_{YX^t\hat{\beta}_n<0} \\ \rightsquigarrow \mathbb{V}(z_\infty + \Gamma) + \mathbb{B}(\beta^*)\end{aligned}$$

under P_n. Note that $\sup_\Gamma \mathbb{G}_n 1_{X^t\beta^*=0} 1_{YX^t\left[\sqrt{n}(\hat{\beta}_n-\beta_n^*)+\Gamma\right]<0}$ is equal to the first term on the right hand side of (7). Hence, $u(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n)$ is the supremum over all local alternatives of the form given in (11). Also taking the supremum over $\Gamma \in \mathbb{R}^p - \{0\}$ we obtain

$$\sup_{\Gamma\in\mathbb{R}^p-\{0\}} \mathbb{V}(z_\infty + \Gamma) + \mathbb{B}(\beta^*) =_{\mathcal{L}} \sup_{u\in\mathbb{R}^p} \mathbb{V}(u) + \mathbb{B}(\beta^*),$$

which is the limiting distribution of $u(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n)$ (see Theorem 3.1). Thus, the ACI can be seen as arising as an asymptotic approximation under local alternatives. This result is formalized below.

Theorem 3.5. *Assume that* $\mathcal{T}_n = \{(X_{ni}, Y_{ni})\}_{i=1}^n$ *is drawn iid from distribution* P_n *for which:*

(B1) $\beta_n^* \triangleq \arg\min_{\beta \in \mathbb{R}^p} P_n L(X, Y, \beta) = \beta^* + \Gamma/\sqrt{n}$ *for some* $\Gamma \in \mathbb{R}^p - \{0\}$,

(B2) if $\mathcal{F}$ *is any uniformly bounded Donsker class and* $\mathbb{G}_n \rightsquigarrow \mathbb{L}$ *in* $l^\infty(\mathcal{F})$ *under* P, *then* $\mathbb{G}_n \rightsquigarrow \mathbb{L}$ *in* $l^\infty(\mathcal{F})$ *under* P_n,

(B3) $\sqrt{n}(\hat{\beta}_n - \beta_n^*) = -H^{-1}\mathbb{G}_n g(X, Y, \beta^*) + o_{P_n}(1)$,

where $\mathbb{G}_n \triangleq \sqrt{n}(\mathbb{P}_n - P_n)$. *Assume (A1)-(A6). Then:*

1. $\mathbb{G}_n 1_{YX^t\hat{\beta}_n < 0} \rightsquigarrow \mathbb{V}(z_\infty + \Gamma) + \mathbb{B}(\beta^*)$

2. $\lim_{n\to\infty} u(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n) =_{\mathcal{L}} \sup_{\eta \in \mathbb{R}^p} \mathbb{V}(z_\infty + \eta) + \mathbb{B}(\beta^*) = \sup_{u \in \mathbb{R}^p} \mathbb{V}(u) + \mathbb{B}(\beta^*)$

under P_n.

Thus the limiting distribution of $u(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n)$ is unchanged under local alternatives and hence might be expected to perform well in small samples. A similar result can be proved for $\ell(\mathbb{G}_n^{(b)}, \hat{\beta}_n^{(b)}, \hat{\Sigma}, a_n)$. This result is underscored by the empirical results in Section 5.

3.4 Choice of Tuning Parameter a_n

Use of the ACI requires the choice of the tuning parameter a_n. We use a simple heuristic for choosing the value of this parameter. The method described here performed well on all of the examples in Section 5. We begin with the presumption that undercoverage is a greater sin than conservatism. Recall that we can view the ACI as a two step procedure where at the first stage we test the null hypothesis $H_0 : x^t\beta^* = 0$ against a two-sided alternative. The test of H_0 used in constructing the ACI rejects when $\frac{(X^t\hat{\beta}_n)^2}{X^t\Sigma X} > \frac{1}{a_n}$. The form of $u(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n)$ in (7) shows that $\frac{1}{a_n}$ too small (e.g. large Type I error) results in too few points being deemed “near the boundary.” Consequently the resulting interval may be too small since the supremum does not affect enough of the training points. Conversely, $\frac{1}{a_n}$ too large (e.g. large Type II error) puts too many points in the region on non-regularity, resulting in an interval that may be too wide because the supremum affects too many of the training points. Given our presumption, controlling Type I error is of primary importance. Let $\gamma \in (0, 1)$. Then let $\frac{1}{a_n} = \frac{1}{\sqrt{n}} \vee \frac{\chi^2_{1-\gamma}}{n}$ and we have for any $x \in \mathbb{R}^p - \{0\}$ and $x^t\beta^* = 0$

$$P\left(\frac{(x^t\hat{\beta}_n)^2}{x^t\Sigma x} > \frac{1}{a_n}\middle| H_0\right) = P\left(\left(\frac{\sqrt{n}(\hat{\beta}_n - \beta^*)^t x}{\sqrt{x^t\Sigma x}}\right)^2 > \frac{n}{a_n}\right) \lesssim \gamma.$$

Thus, the suggested a_n controls the Type I error to be no more than γ. Moreover, it is clear from the above display that the Type I error decreases to zero as n tends to infinity. In all of the experiments in this paper we choose, rather arbitrarily, to use $\gamma = .005$. Simulations results, given in Table 5 of the online supplement, show that the performance (measured in terms of width and coverage) of the ACI appears to be insensitive to choices of γ in the range .001 to .01 for a sample size of around 30. For larger sample sizes, the choice of γ is unimportant since $\sqrt{n} > \chi^2_{1-\gamma}$ except for extremely small values of γ.

4 Computation

To implement the ACI we need to calculate, for each bootstrap sample, the supremum and infimum in $u(\mathbb{G}_n^{(b)}, \hat{\beta}_n^{(b)}, \hat{\Sigma}_n, a_n)$, and $l(\mathbb{G}_n^{(b)}, \hat{\beta}_n^{(b)}, \hat{\Sigma}_n, a_n)$ respectively. The required optimization, as stated, is a Mixed Integer Program (MIP) because of the discrete nature of the indicator function. In this section, we develop a convex relaxation that can be solved in polynomial time. The details for the infimum are provided below; a similar approach is used to find the supremum by writing $1_{z<0} = 1 - 1_{z\geq 0}$ and using the relationship: $\sup_z g(z) = -inf_z - g(z)$. Let $(m_{n1}, m_{n2}, \ldots, m_{nn})$ be a realization of the bootstrap weights

$(M_{n1}, M_{n2}, \ldots, M_{nn}) \sim Multinomial(n, \frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n})$. For each such realization, construction of the infemum in the ACI requires computing

$$\inf_{u \in \mathbb{R}^p} \sum_{i \in N_n^{(b)}} (m_{ni} - 1) 1_{y_i x_i^t u < 0}, \tag{12}$$

where $N_n^{(b)} = \{i : \frac{(x_i^t \hat{\beta}_n^{(b)})^2}{x_i^t \hat{\Sigma}_n x_i} \leq \frac{1}{a_n}\}$. In this form, the optimization is clearly seen to be an MIP. Reliably solving an MIP requires the use specialized software (we use CPLEX) and quickly becomes computationally burdensome as the size of the problem grows. The following convex relaxation of (12) is (i) computationally efficient requiring roughly the same amount of computation as fitting a linear SVM and (ii) can be solved without specialized software (e.g. R or matlab).

As the initial step write

$$\sum_{i \in N_n^{(b)}} (m_{ni} - 1) 1_{y_i x_i^t u < 0} = \sum_{i \in N_n^{(b)}} m_{ni} 1_{y_i x_i^t u < 0} + \sum_{i \in N_n^{(b)}} (-1_{y_i x_i^t u < 0}).$$

Then replace the indicator function $1_{y_i x_i^t u < 0}$ with convex surrogate and upper bound $(1 - y_i x_i^t u)_+$ where $(z)_+$ denotes the positive part of z. Similarly, replace the function $-1_{y_i x^t u < 0}$ with convex surrogate and upper bound $(1 + y_i x_i^t u)_+ - 1$. The indicator functions and their respective surrogates are shown in Figure 2. The relaxed optimization problem is then

$$\inf_{u \in \mathbb{R}^p} \sum_{i \in N_n^{(b)}} \left[m_{ni}(1 - y_i x_i^t u)_+ + (1 + y_i x_i^t u)_+ \right] \tag{13}$$

where the -1 in the relaxation of $-1_{y_i x_i^t u < 0}$ has been omitted since it does not depend on u. The optimization problem in (13) can be cast as a linear program and hence solved in polynomial time. See the next section for an empirical comparison of the relaxed and MIP solutions to (12).

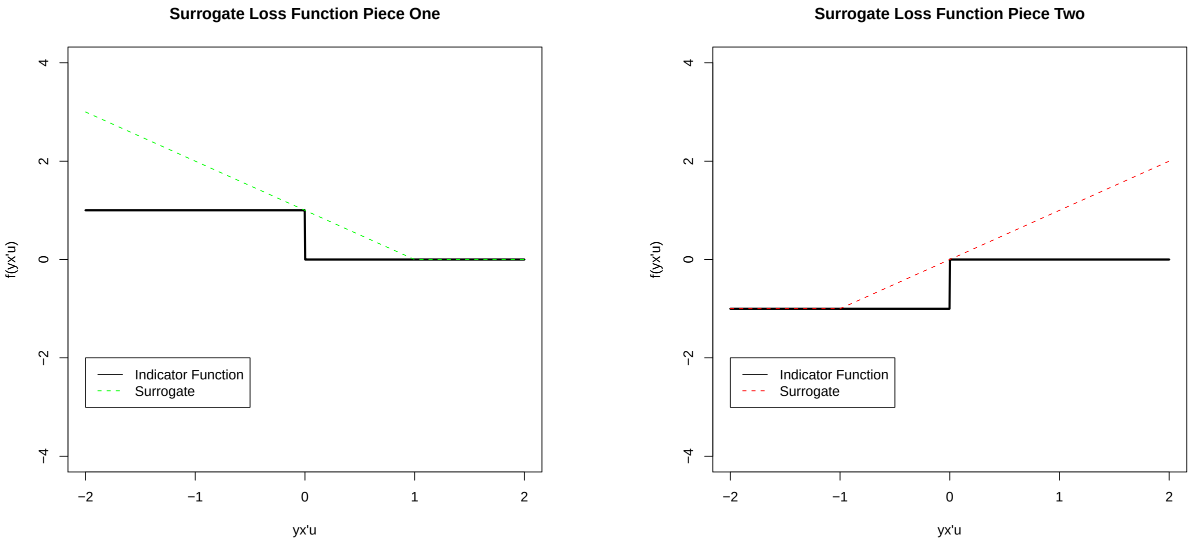

Figure 2: Relaxation of the indicator functions. Left panel: indicator function $1_{yx^tu<0}$ replaced with convex surrogate $(1 - yx^tu)_+$. Right panel: indicator function $-1_{yx^tu<0}$ replaced with convex surrogate $(1 + yx^tu)_+ - 1$.

5 Empirical study

In this section we compare solution quality between the relaxed and MIP solutions to (12); as will be seen the relaxed solution to (12) can be computed much more quickly while little is lost in terms of solution quality. Next using the relaxed solution to (12) the empirical performance of the ACI is compared with two recent methods proposed in the literature. Ten data sets are used in these comparisons; three are simulated and the remaining seven data sets are taken from the UCI machine learning repository (www.ics.uci.edu/∼mlearn/MLRepository.html) and thus the true generative model is unknown. In this case, the empirical distribution function of the data set is treated as the generative model. A summary of the data sets are given in Table 2.

To assess the difference in solution quality between the relaxed and MIP solutions to (12) we perform the following procedure for each of the 10 examples listed in Table 2. We generate 1000 training sets of size $n = 30$, and for each training set we compute 1000 bootstrap resamples. For each resample we compute (12) exactly using the MIP and approximately using the convex relaxation described above. Here we illustrate the results when the loss function used to construct $\hat{\beta}_n$ and $\hat{\beta}_n^{(b)}$ is chosen to be $L(X, Y, \beta) = (1 - YX^t\beta)^2$. Let $\theta_{MIP}^{(t)(b)}$ and $\theta_{REL}^{(t)(b)}$ denote the MIP and relaxed solution to (12) for the b^{th} bootstrap resample of the t^{th} training set. Table 1 reports the 50, 75, 95, and 99 percentiles of $\frac{1}{n}\left(\theta_{MIP}^{(t)(b)} - \theta_{REL}^{(t)(b)}\right)$ for each example. Notice that for each example we considered, the relaxed and MIP solutions agree exactly on more than half of the resampled pairs. Moreover, on more than 95 percent of the resampled pairs, we observe that $\frac{1}{n}\left(\theta_{MIP}^{(t)(b)} - \theta_{REL}^{(t)(b)}\right) \leq \frac{1}{n}$, implying that the two solutions differed by at most the activation of a single indicator function. Table 1 also reports the estimated coverage of confidence sets constructed using the MIP and relaxed formulations. For each of the 10 data sets, estimated coverage using the two methods is not significantly different. The final bit of information in Table 1 regards computation time. The last two columns report the average time in seconds that it takes to construct a single confidence interval using the MIP and relaxed formulations. Computations were performed using a 3.06 GHz intel processor with 4 GB 1067 MHz DDR3. It is clear that even in the $n = 30$ case significant computational gain can be made by using the relaxed formulation. However, this gain becomes more pronounced as sample size increases. Figure 3 compares the computation time for the ThreePt data set (this data set is decribed in Laber and Murphy 2009) as a function of sample size using squared error loss. As claimed, the computation time for the relaxed construction scales much more efficiently than the MIP formulation. In the examples presented in the next section we use the convex relaxation to compute the confidence interval.

5.1 Competing methods

As competitors we consider a repeated-split normal approximation suggested by (Yang 2006) and the recently proposed Bootstrap Case Cross-Validated Percentile with Bias Reduction

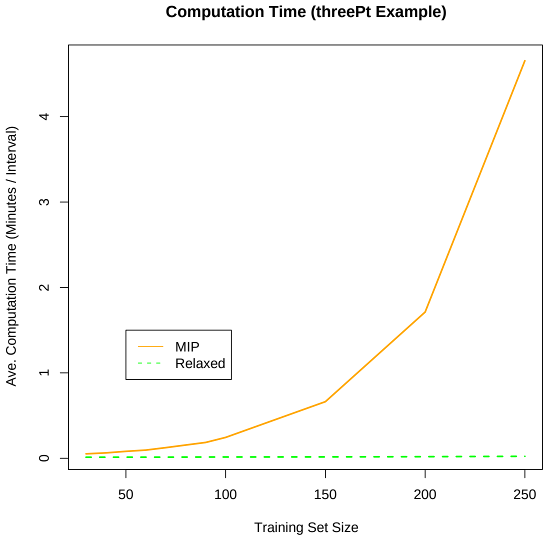

Figure 3: Computation time for MIP and relaxed construction of ACI using the ThreePt data set and squared error loss.

	Coverage		Difference in width				Computation time	
Data Set	Relaxed	MIP	$p_{.99}$	$p_{.95}$	$p_{.75}$	$p_{.5}$	Relaxed	MIP
ThreePt	.948	.948	.0334	0.00	0.00	0.00	.734	3.11
Magic	.944	.945	.0334	.0334	0.00	0.00	1.24	1.94
Mam.	.957	.958	.0334	0.00	0.00	0.00	.904	1.88
Ion.	.954	.954	.0334	0.00	0.00	0.00	1.33	3.06
Donut.	.967	.968	.0667	.0334	.0334	0.00	.917	2.94
Bal.	.969	.969	0.00	0.00	0.00	0.00	.977	1.69
Liver	.956	.956	.0333	.0333	0.00	0.00	1.61	2.50
Spam	.984	.987	.0333	.0333	0.00	0.00	1.54	3.01
Quad	.959	.962	.0333	0.00	0.00	0.00	.983	1.37
Heart	.960	.961	.0333	0.00	0.00	0.00	1.06	3.27

Table 1: Comparison of MIP and relaxed versions of the ACI. For each data set the table was constructed using 1000 training sets each with 1000 bootstrap iterations for a total of 1,000,000 computations of the optimization problem given in (12).

(BCCVP-BR) method of (Jiang 2008). These methods represent the best we could find in terms of consistent coverage. Both methods substantially outperform standard approaches like the bootstrap and normal approximation which are discussed in Section 2. To provide a baseline for comparison, the performance of the Centered Percentile Bootstrap (CPB) is included in the online supplement.

Briefly, Yang's method repeatedly partitions the training data $\mathcal{T}$ into two equal halves $\mathcal{T}^L$ and $\mathcal{T}^V$. A classifier is trained on $\mathcal{T}^L$ and then evaluated on $\mathcal{T}^V$. The mean and variance of the number of misclassified points in $\mathcal{T}^V$ is recorded. This mean and variance are then aggregated and used in a normal approximation. Jiang's method can be roughly described as leave one out cross validation with bootstrap resamples. However, since a bootstrap resample can have multiple copies of a single training example, leave one out cross-validation will no longer have disjoint training and testing sets. Instead, for each unique training example (x_i, y_i) the bootstrap resample is partitioned into two sets, one with all copies of (x_i, y_i) call this $\mathcal{V}$, and the second contains the remainder of the resample call this $\mathcal{L}$. The classifier is trained on $\mathcal{L}$ and evaluated on $\mathcal{V}$. The average error over all sets $\mathcal{V}$ is recorded within each bootstrap resample and the percentiles form the endpoints of a confidence interval. As a final step Jiang provides a bias correction. A full description of these methods can be found in the referenced works. While these methods are intuitive, they lack theoretical justification. Yang's method was developed for use with a hold-out set; when such a hold-out set does not exist, the method is inconsistent. Jiang offers no justification other than intuition.

5.2 Results

We examine the performance of the ACI and competing methods using the following three metrics (i) coverage (ii) interval width and (iii) computational expense. These metrics are recorded using ten data sets, three sample sizes, and three loss functions. Three of the examples use simulated datasets and hence the test error can be computed exactly. The remaining seven data sets are taken from the UCI machine learning repository (www.ics.uci.edu/∼mlearn/MLRepository.html) and thus the true generative model is unknown. In this case, the empirical distribution function of the data set is treated as the generative model. Results using squared error loss are listed here while the results using binomial deviance and ridged hinge loss (support vector machines) are given in the online supplement. A summary of the data sets are given in Table 2.

Coverage results for squared error loss are given in Table 3. The adaptive confidence interval is the only method to attain at least nominal coverage on all ten test sets. Yang's method is either extremely conservative or anti-conservative. Jiang's interval attains the nominal coverage on eight of ten data sets in the $n = 30$ case and nine of ten data sets for larger sample sizes. Table 4 shows the width of the constructed confidence intervals. When $n = 30$ the ACI is smallest in width for eight of the ten data sets. For larger sample sizes Jiang's method and the ACI display comparable widths; Yang's method is always the widest. Another important factor is computation time. Table 5 shows the average amount of time required in seconds to construct a single confidence interval. All methods used 1000 resamples. That is, 1000 bootstrap resamples for the ACI and Jiang's method, and 1000 repeated splits for Yang's method. Table 5 shows that Yang's method is the most computationally efficient. However, it is also clear that Jiang's method is significantly slower than the ACI for moderate sample sizes. For the Magic data set Jiang's method takes more than 30 times longer than the ACI. It is most important, however, to notice the trend in computation time across sample sizes. Computation time for Yang's method and the ACI grow slowly with sample size while the computational cost of Jiang's method increases much more quickly. The reason for this is that Jiang's method performs leave-one-out cross validation for each bootstrap resample thus increasing the computation time by a factor of n. Results for ridged hinge loss and binomial deviance loss are similar and can be found in the technical report (Laber and Murphy, 2010).

Name	Features	Source	$\mathbb{E}\tau(\hat{c})$ (SE)	$\mathbb{E}\tau(\hat{c})$ (BD)	$\mathbb{E}\tau(\hat{c})$ (SVM)
ThreePt	2	Simulated	.500	.500	.500
Quad	3	Simulated	.0997	.109	.101
Donut	3	Simulated	.235	.249	.232
Magic	11	UCI	.264	.231	.252
Mam.	6	UCI	.192	.190	.203
Ion.	9	UCI	.151	.147	.149
Bal.	5	UCI	.054	.050	.061
Liver	7	UCI	.342	.342	.334
Spam	10	UCI	.190	.183	.181
Heart	9	UCI	.167	.173	.174

Table 2: Test data sets used to evaluate confidence interval performance. The last three columns record the average test error for a linear classifier trained using a training set of size $n = 100$ and loss function: squared error loss (SE), binomial deviance (BD), and ridged hinge loss (SVM).

Sample Size	$n = 30$			$n = 100$			$n = 250$		
Data Set / Method	ACI	Yang	Jiang	ACI	Yang	Jiang	ACI	Yang	Jiang
ThreePt	.948	.930*	.863*	.937	.537*	.925*	.935	.387*	.930*
Magic	.944	.996*	.979*	.973*	.991*	.969*	.962	.996*	.974*
Mam.	.957	.989*	.966	.937	.996*	.964	.960	.995*	.968
Ion.	.941	.996*	.972*	.961	.992*	.964	.952	.996*	.949
Donut	.965	.967	.908*	.970*	.866*	.974*	.974*	.895*	.988*
Bal.	.976*	.989*	.966	.962	.995*	.969*	.946	.991*	.963
Liver	.956	.997*	.970*	.963	.992*	.966	.971*	.996*	.984*
Spam	.984*	.998*	.975*	.967	.996*	.967	.979*	.996*	.958
Quad	.959	.983*	.945	.957	.989*	.938	.965	.999*	.940
Heart	.960	.995*	.976*	.949	.991*	.979*	.971*	.989*	.974*

Table 3: Coverage comparison between ACI, Yang's CV and Jiang's $BCCVP - BR$ for squared error loss, target coverage is .950. Coverage is starred if observed coverage is significantly different from .950 at .01 level.

6 Discussion

Many statistical procedures in use today are justified by a combination of asymptotic approximations and high quality simulation performance. As exemplified here, the choice of asymptotic framework may be crucial in obtaining reliably good performance in small sam-

Sample Size	$n = 30$			$n = 100$			$n = 250$		
Data Set / Method	ACI	Yang	Jiang	ACI	Yang	Jiang	ACI	Yang	Jiang
ThreePt	.385*			.198*			.193*		
Magic	.498*	.528	.501	.238	.257	.214*	.125	.157	.122*
Mam.	.374*	.456	.383	.191	.226	.178*	.112	.140	.105*
Ion.	.313*	.466	.388	.175	.213	.172*	.103	.127	.100*
Donut	.424*	.483		.217*	.258		.123*		.201
Bal.	.217*	.350	.232	.101*	.138	.103	.0623	.0772	.0620*
Liver	.534	.527	.500*	.262	.274	.241*	.152	.172	.143*
Spam	.428	.496	.418*	.219	.229	.184*	.125	.140	.108*
Quad	.246*	.360	.267	.142*	.171	.144	.0811*	.104	.0885
Heart	.367*	.476	.404	.184*	.219	.184*	.106*	.132	.110

Table 4: Comparison of interval width between ACI, Yang's CV and Jiang's $BCCVP-BR$ for squared error loss. Smallest observed width is starred. Examples where at least the nominal coverage was not attained are omitted.

Sample Size	$n = 30$			$n = 100$			$n = 250$		
Data Set / Method	ACI	Yang	Jiang	ACI	Yang	Jiang	ACI	Yang	Jiang
ThreePt	.734			.762			1.37		
Magic	1.24	.0392	1.59	1.40	.0834	11.1	1.90	0.178	60.66
Mam.	1.37	.0185	.697	6.03	.0383	5.52	12.8	.0800	26.3
Ion.	2.13	.0331	1.32	6.42	.0702	10.0	16.7	.147	52.62
Donut	2.00	.00930		4.33		2.16	11.6		10.84
Bal.	.977	.0160	.575	1.05	.0315	3.50	1.23	.0660	20.9
Liver	1.16	.0222	.859	1.44	.0461	6.25	1.78	.0978	33.7
Spam	1.38	.0348	1.37	1.53	.0744	10.5	1.72	.159	57.9
Quad	.983	.00918	.125	1.11	.0191	1.43	1.24	.0398	6.96
Heart	1.06	.0317	1.25	1.15	.0660	8.00	1.42	.139	23.6

Table 5: Comparison of computation time (in seconds) between ACI, Yang's CV and Jiang's $BCCVP - BR$ for squared error loss. Examples where at least the nominal coverage was not attained are omitted.

ples. In this paper a non-regular asymptotic framework in which the limiting distribution of the test error changes abruptly with changes in the true, underlying data generating distribution is used to develop a confidence interval. In particular, asymptotic non-regularity occurs due to the non-smooth test error in connection with particular combinations of β^* values and the X distribution. It is common practice to "eliminate" this asymptotic non-regularity by assuming that these problematic combinations of β^* values and the X distribution cannot occur. However, small samples are unable to precisely discriminate between settings that are *close to* the problematic β^* values/X distribution from settings in which the β^* values/X distribution are *exactly* problematic. As a result, asymptotic approximations that depend on assuming away these problematic settings can be of poor quality; this is the case here.

The validity of proposed adaptive confidence interval presented here does not depend on assuming away problematic scenarios; instead the ACI detects and then accommodates settings that are sufficiently close to the problematic β^* values/X distribution. In this sense the ACI adapts to the non-smoothness in the test error. Specifically, in settings in which standard asymptotic procedures fail, the ACI provides asymptotically valid, albeit conservative, confidence intervals. Moreover, the ACI delivers exact coverage if either (i) the model space is correct or (ii) a margin condition holds. Practically, this means that in a setting where standard asymptotic procedures (e.g. the bootstrap) are applicable, the ACI is asymptotically equivalent to these methods. Experimental performance of the ACI is also quite promising. On a suite of 10 examples, three loss functions and three classification algorithms, the ACI delivered nominal coverage. In addition, the ACI generally had a smaller length than competing methods. The ACI can be computed efficiently with algorithms scaling polynomially in dimension and sample size.

Two important extensions of the ACI are: first, to extend the ACI to construct valid confidence intervals for the difference in test error between two linear classifiers and, second, to extend these ideas to the setting in which the number of features is comparable or larger than the sample size. The former extension is straightforward and can be achieved by enlarging the set over which the supremum is taken in (7) to include the points on the classification boundaries of both classifiers. The latter is more difficult. In the estimation of classifiers in the $p >> n$ setting, it is important to avoid overfitting. A typical approach to reduce the amount of overfitting is regularization which effectively reduces the space of available classifiers to choose from. Similarly, the supremum in (7) must be taken over a restricted set of classifiers to avoid being unnecessarily wide. Extending the theory and computation to this setting is left to another paper.

References

Anthony, M. M. and Bartlett, P. (1999), *Learning in Neural Networks: Theoretical Foundations*, New York, NY, USA: Cambridge University Press.

Bartlett, P., Jordan, M., and McAuliffe, J. (2005), "Convexity, classification, and risk bounds," *Journal of the American Statistician*, 101, 138–156.

Bickel, P., Klaassen, A., Ritov, Y., and Wellner, J. (1993), *Efficient and adaptive inference in semi-parametric models*, Johns Hopkins University Press, Baltimore.

Bose, A. and Chatterjee, S. (2000), "Generalized bootstrap for estimators of minimizers of convex functionals," Tech. rep., Indian Statistical Institute.

— (2003), "Generalized bootstrap for estimators of minimizers of convex functions," *Journal of Statistical Planning and Inference*, 117, 225 – 239.

Cheng, X. (2008), "Robust Confidence Intervals in Nonlinear Regression Under Weak Identification," *Job Market Paper*.

Chernick, M., Murthy, V., and Nealy, C. (1985), "Application of Bootstrap and Other Resampling Techniques: Evaluation of Classifier Performance," *PRL*, 3, 167–178.

Chung, H.-C. and Han, C.-P. (2009), "Conditional confidence intervals for classification error rate," *Computational Statistics and Data Analysis*, 53, 4358–4369.

Donald, D. W. (2001), “Testing when a parameter is on the Boundary of the Maintained Hypothesis,” *Econometrica*, 69, 683–734.

Efron, B. (1983), “Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation,” *Journal of the American Statistical Association*, 78, 316–331.

Efron, B. and Tibshirani, R. (1995), “Cross-Validation and the Bootstrap: Estimating the Error Rate of a Prediction Rule,” Tech. Rep. 172, Stanford.

— (1997), “Improvements on Cross-Validation: The .632+ Bootstrap Method,” *Journal of the American Statistical Association*, 92, 548–560.

Haberman, S. (1989), “Concavity and Estimation,” *Annals of Statistics*, 17, 1631–1661.

Hastie, T., Tibshirani, R., and Friedman, J. (2009), *The Elements of Statistical Learning*, Springer Series in Statistics, New York, NY, USA: Springer New York Inc.

Isaksson, A., Wallman, M., Gransson, H., and Gustafsson, M. (2008), “Cross-validation and bootstrapping are unreliable in small sample classification,” *Pattern Recognition Letters*, 29, 1960 – 1965.

Jiang, W., Varma, S., and Simon, R. (2008), “Calculating confidence intervals for prediction error in microarray classification using resampling,” *Statistical Applications in Genetics and Molecular Biology*, 7.

Kohavi, R. (1995), “A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection,” in *IJCAI*, Morgan Kaufmann, pp. 1137–1145.

Kosorok, M. (2008), *Introduction to empirical processes and semiparametric inference*, Springer Verlag.

Krzanowski, W. and Hand, D. (1985), "Assessing Error Rate Estimators: The Leave-One-Out Method Reconsidered," *PRL*, 3, 167–178.

Laber, E. B. and Murphy, S. A. (2008), "Small Sample Inference for Generalization Error in Classification Using the CUD Bound," in *Proceedings of the Proceedings of the Twenty-Fourth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-08)*, Corvallis, Oregon: AUAI Press, pp. 357–365.

— (2009), "Adaptive Confidence Intervals for the Test Error in Classification," Tech. Rep. 497, University of Michigan.

Niemiro, W. (1992), "Asymptotics for M-Estimators defined by convex minimization," *Annals of Statistics*, 20, 1514–1533.

Schiavo, R. A. and Hand, D. (2000), "Ten More Years of Error Rate Research," *International Statistical Review*, 68, 295–310.

Van der Vaart, A. and Wellner, J. (1996), *Weak convergence and empirical processes: with applications to statistics*, Springer Verlag.

Xie, M., Singh, K., and Zhang, C.-H. (2009), "Confidence Intervals for Population Ranks in the Presence of Ties and Near Ties," *Journal of the American Statistical Association*, 104, 775–788.

Yang, Y. (2006), "Comparing Learning Methods for Classification," *Statistica Sinica*, 16, 635–657.

Zhang, P. (1995), "APE and Models for Categorical Panel Data," *Scandinavian Journal of Statistics*, 83–94.