Published in final edited form as:

J Am Stat Assoc. 2011 July 1; 106(495): 920–924. doi:10.1198/jasa.2011.tm11320.

Comment on “Adaptive Confidence Intervals for the Test Error in Classification”

Yair Goldberg and

postdoctoral scholar in the Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599 (ygoldeber@bios.unc.edu)

Michael R. Kosorok

Professor and Chair of the Department of Biostatistics and Professor of the Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599 (kosorok@unc.edu)

Abstract

Inspired by the non-regular framework studied in Laber and Murphy (2011), we propose a family of adaptive classifiers. We discuss briefly their asymptotic properties and show that under the non-regular framework these classifiers have an “oracle property,” and consequently have smaller asymptotic variance and smaller asymptotic test error variance than those of the original classifier. We also show that confidence intervals for the test error of the adaptive classifiers, based on either normal approximation or centered percentile bootstrap, are consistent.

1 Introduction

Laber and Murphy (2011) use an interesting non-regular framework to examine the behavior of the test error for a wide class of linear classifiers. The non-regular framework highlights why existing confidence intervals do not work well for small-sized samples. Using this new understanding, the authors introduce a novel adaptive confidence interval that outperforms existing confidence intervals. The use of the chosen non-regular framework not only paves the way for the adaptive confidence interval, but also provides an important new tool for understanding and development of classifiers.

The authors discuss a general linear classifier that is obtained as a minimizer of the empirical expectation of some convex criterion function. Even within this seemingly simple framework, estimation of the test error proves to be a difficult task. The authors therefore consider the case in which the optimal decision boundary has positive probability, and the distribution of the labels on the optimal decision boundary is not even. In this case, the estimated test error converges to a non-degenerate distribution. The authors refer to this scenario as the *non-regular framework*.

Using the non-regular framework, the authors show that the scaled and centered test error can be decomposed into two components, corresponding to points that are on the optimal decision boundary and points that are not (see Laber and Murphy 2011, Eq. 6). Inspired by this decomposition, the authors develop an adaptive confidence interval. Moreover, using this decomposition, the authors show that confidence intervals based on either normal approximation or the centered percentile bootstrap are not consistent. These results

seemingly imply that test-error estimation is more difficult within the non-regular framework.

In the following, we present a family of “adaptive classifiers” that are obtained by adding an adaptive penalization term to the original minimization problem discussed by Laber and Murphy (2011). We show that within the regular framework, the effect of the penalization term tends to zero, namely, the adaptive classifier has the same asymptotic behavior as that of the minimizer of the original problem. Within the non-regular framework, these adaptive classifiers perform (asymptotically) as well as if the optimal decision boundary were known. Moreover, these classifiers have smaller asymptotic variance and smaller asymptotic test error variance than those of the original classifier. Finally, we show that within both regular and non-regular frameworks, confidence intervals for the test error of the adaptive classifiers, based on either normal approximation or centered percentile bootstrap, are consistent.

We first present the proposed penalized criterion function and then present the proposed adaptive classifier and its asymptotic properties, including an intriguing oracle result. A brief discussion then follows, and proofs are given afterwards.

2 Penalizing the Criterion Function

In the following we introduce linear classifiers based on minimizing the empirical expectation of some convex criterion function plus a novel penalization term. We follow the notation of Laber and Murphy (2011) whenever possible.

Let the training set be a set of n i.i.d. pairs $\mathcal{T}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$. Let $L(X, Y, \beta)$ be a convex loss function in β almost surely. Denote $\hat{\beta}_n = \operatorname{argmin}_{\beta} \mathbb{P}_n L(X, Y, \beta)$ and $\beta^* = \operatorname{argmin}_{\beta} E[L(X, Y, \beta)]$. The test error is defined as $E[Y X' \hat{\beta}_n] < 0$.

As shown by the authors (Laber and Murphy 2011, Section 2), the structure of the test error is such that small perturbations of the classification vector may alter the test error drastically. This is especially true when the probability of points being close to the decision boundary is large. Since $\sqrt{n}(\hat{\beta}_n - \beta^*) = O_p(1)$, such perturbations are expected.

Our goal is thus to force the classifier towards the optimal decision boundary, i.e., to find a classifier with smaller angle with β^* . This can be done by penalizing the angle between the classifier vector and points that are on or close to the boundary. However, since the optimal boundary is not known, one needs to estimate it. Estimating which points are close to or on the boundary can be done using hypothesis testing as done by the authors (see Laber and Murphy 2011, Section 3.1). Here we take a different approach that is based on adaptive weights. Define

$$Q_n(\beta) = \sum_{i=1}^n L(x_i, y_i, \beta) + \frac{\lambda_n}{n} \sum_{i=1}^n \hat{w}_{ni} |\beta' x_i| = \mathbb{P}_n \left(nL(X, Y, \beta) + \lambda_n \hat{W} |\beta' X| \right), \quad (1)$$

for some $\lambda_n > 0$ and weights $\hat{w}_{ni}$, and where $\hat{W}$ will be defined shortly. Let $\tilde{\beta}_n = \operatorname{argmin} Q_n(\beta)$ be the obtained classifier. Depending on the choice of weights, the classifier $\tilde{\beta}_n$ can have oracle properties, as will be discussed in the next section.

In the remainder of the section we discuss the choice of weights $\hat{w}_{ni}$. Since we would like to penalize points that are close to or on the boundary we would like to demand that $\hat{w}_{ni}$ be large for such points. On the other hand, we are not interested in penalizing points that are not close to the boundary, hence we would like $\hat{w}_{ni} \rightarrow 0$ for such points.

We Define the weights $\widehat{w}_{ni} = n^{1/2+\alpha}(1 - n^\alpha \widehat{\beta}_n' X_i)_+$ for some fixed $\alpha \in (0, 1/2)$, and where $(a)_+ = \max\{a, 0\}$. This implies that $\widehat{W} = \widehat{W}(X) = n^{1/2+\alpha} \left(1 - n^\alpha \widehat{\beta}_n' X\right)_+$. Note that the weight $\widehat{w}_{ni}$ is large whenever $\widehat{\beta}_n' X_i$ is small, i.e., when X_i is close to the decision boundary of $\widehat{\beta}_n$. Since $\widehat{\beta}_n - \beta^* = O_p(n^{-1/2})$, the weight is large whenever X_i is close to the optimal decision boundary. Moreover, when X_i is far from the optimal decision boundary, the weight is small or zero.

More formally, Define $M_n = \{i : \beta^{*'} X_i > 0\}$ and let $M_n^c = \{i : \beta^{*'} X_i \leq 0\}$. We need the following assumptions:

- (B1) X is bounded, i.e., there is a constant M such that $\|X\| \leq M$ almost surely.
- (B2) $P(|X' \beta^*| < t | X' \beta^* > 0) \geq Ct$ for some constant C and all t small enough.
- (B3) $\text{Var}(X)$ is positive definite.

We remark that Assumption (B1) can be relaxed with the price of complicating the proofs and that Assumption (B2) holds whenever $X|X' \beta^* > 0$ has bounded density.

Lemma 2.1. Assume (B1)–(B3). Then

$$\max_{i \in M_n} \frac{\widehat{w}_{ni}}{\sqrt{n}} = o_p(1), \quad (2)$$

and

$$\min_{i \in M_n^c} \frac{\widehat{w}_{ni}}{\sqrt{n}} \rightarrow_p \infty. \quad (3)$$

Proof of Lemma 2.1. First note that by Assumption (B2), for all $0 < \alpha < 1/2$,

$$P\left(\min_{i \in M_n} |\beta^{*'} X_i| \leq \frac{t}{n^\alpha}\right) = 1 - \prod_{i \in M_n} \left(1 - P\left(|\beta^{*'} X_i| \leq \frac{t}{n^\alpha}\right)\right) \leq 1 - \left(1 - \frac{Ct}{n^\alpha}\right)^{\#(M_n)}, \quad (4)$$

where $\#(M_n)$ is the cardinality of the set M_n . Write $1 - q = \lim_{n \rightarrow \infty} \#(M_n)/n > 0$, and note that

$$\lim_{n \rightarrow \infty} 1 - \left(1 - \frac{Ct}{n^\alpha}\right)^{\#(M_n)} = \lim_{n \rightarrow \infty} 1 - \exp\{-Ct(1 - q)n^{1-\alpha}\} = 0. \quad (5)$$

We are now ready to prove (2).

$$P\left(\max_{i \in M_n} \frac{\widehat{w}_{ni}}{\sqrt{n}} > 0\right) = P\left(\max_{i \in M_n} (1 - n^\alpha \widehat{\beta}_n' X_i) > 0\right) = P\left(\min_{i \in M_n} |(\widehat{\beta}_n - \beta^*)' X_i + \beta^{*'} X_i| < n^{-\alpha}\right) \leq P\left(\min_{i \in M_n} |\beta^{*'} X_i| < n^{-\alpha} + M \|\widehat{\beta}_n - \beta^*\|\right). \quad (6)$$

Here we used Assumption (B1) for the last inequality. Since $(\widehat{\beta}_n - \beta^*) = O_p(n^{-1/2}) = o_p(n^{-\alpha})$, for all n large enough we can bound (6) using (4) with $t = 2$. Applying (5) we obtain that (2) holds.

To see that (3) holds, note that

$$\min_{i \in M_n^c} \widehat{w}_{ni} / \sqrt{n} = n^\alpha (1 - \max_{i \in M_n^c} n^\alpha |(\widehat{\beta}_n - \beta^*)' X_i|) \geq n^\alpha (1 - M \|\widehat{\beta}_n - \beta^*\|).$$

Recall that $\hat{\beta}_n - \beta^* = O_p(n^{-1/2}) = o_p(n^{-\alpha})$. Hence $P(n^\alpha \|\hat{\beta}_n - \beta^*\| < 1/2) \rightarrow_p 1$. This implies that $n^\alpha(1 - n^\alpha \|\hat{\beta}_n - \beta^*\|) \rightarrow_p \infty$ and (3) holds.

3 The Adaptive Classifier and its Asymptotic Properties

In the previous section we defined the adaptive classifier $\tilde{\beta}_n$ as the minimizer of $Q_n(\beta)$ (see 1). In this section we study the asymptotic properties of $\tilde{\beta}_n$.

Define the non-regular condition as

$$(NR) P(X' \beta^* = 0) = q > 0.$$

It follows from Lemma 2.1 that when condition (NR) does not hold, the penalization term has small influence on the minimizer of (1). However, as we prove below, when condition (NR) does hold, the penalization term forces the classifier to be in the linear span of the optimal classifier β^* with probability that tends toward one. We refer to this phenomenon as the *oracle property* of the classifier. There are three main consequences of the oracle property. First, the quantity $\sqrt{n}(\tilde{\beta}_n - \beta^*)$ has smaller asymptotic variance than $\sqrt{n}(\hat{\beta}_n - \beta^*)$. Second, the test error has smaller asymptotic variance than the one based on $\hat{\beta}_n$ due to the vanishing contribution to the variance of the test error from the points on the optimal decision boundary. Finally, the confidence intervals that are based on the normal approximation and centered percentile bootstrap are consistent.

Remark 3.1. The oracle property discussed here is different from the oracle property in the context of variable selection (see, for example, Fan and Li 2001; Zou 2006). In the context of variable selection the oracle property means that if there are zero components, they are estimated as zero with probability tending to 1, and the nonzero components are estimated as well as if the correct submodel were known. Here, when the non-regular condition holds, the optimal decision boundary obtains with probability tending to 1, and the classifier is estimated as if the correct subspace $\text{span}(\beta^*)$ were known.

We start by discussing the properties of $\tilde{\beta}_n$. Formally, let $V = \text{span}(\beta^*)$, $V^\perp$ be the orthocomplement of V in $\mathbb{R}^p$, and let $\mathbf{v} = \beta^* / \|\beta^*\|$. Let $P_V \equiv \mathbf{v}\mathbf{v}'$ and $P_{V^\perp}$ be the projection matrices on V and $V^\perp$, respectively.

Theorem 3.2. Assume Assumptions (A1)–(A6) of Laber and Murphy (2011) and Assumptions (B1)–(B3). Let λ_n be a sequence bounded away from zero and infinity. Then $\tilde{\beta}_n \rightarrow_{a.s.} \beta^*$, and

OLD VERSION

$$\sqrt{n}(\tilde{\beta}_n - \beta^*) \rightarrow_d \begin{cases} N(\mathbf{0}, P_V H^{-1} P_V \Omega P_V H^{-1} P_V) & \text{when condition (NR) holds} \\ N(\mathbf{0}, H^{-1} \Omega H^{-1}) & \text{when condition (NR) does not hold} \end{cases}.$$

New Version

$$\sqrt{n}(\tilde{\beta}_n - \beta^*) \rightarrow_d \begin{cases} N(\mathbf{0}, (\mathbf{v}' H \mathbf{v})^{-2} P_V \Omega P_V) & \text{when condition (NR) holds} \\ N(\mathbf{0}, H^{-1} \Omega H^{-1}) & \text{when condition (NR) does not hold} \end{cases}. \quad (7)$$

Moreover, when condition (NR) holds the classifier is oracle, i.e., with probability tending to 1

$$\tilde{\beta}_n \in \text{span}(\beta^*). \quad (8)$$

See the proof given in the Appendix.

We are now ready to discuss the test error of the adaptive classifier. Denote

$$\begin{aligned}\sigma_1^2 &= E[(1_{X'\beta^* \neq 0} 1_{YX'\beta^* < 0} - E 1_{X'\beta^* \neq 0} 1_{YX'\beta^* < 0})^2], \\ \widehat{\sigma}_1^2 &= \mathbb{P}_n 1_{YX'\beta < 0} (1 - \mathbb{P}_n 1_{YX'\beta < 0}).\end{aligned}$$

Then an asymptotic $(1 - \gamma)100\%$ confidence interval for the test error $\tau(\tilde{c})$ based on the

normal approximation is $\widehat{\tau}(\tilde{c}) \pm z_{1-\gamma/2} \frac{\widehat{\sigma}_1}{\sqrt{n}}$, where τ is defined as in Laber and Murphy (2011, Section 2), and $\tilde{c}$ is defined similarly to $\hat{c}$.

An alternative way to construct the confidence interval can be based on bootstrapping. More specially, let $\mathbb{P}_n^{(b)}$ be the bootstrap empirical measure with multinomial weights (see Laber and Murphy 2011, Section 2). Let $\widehat{\beta}_n^{(b)} = \text{argmin}_{\beta} \mathbb{P}_n^{(b)} L(X, Y, \beta)$. Define

$\widehat{w}_{ni}^{(b)} = n^{1/2+\alpha} (1 - n^\alpha \widehat{\beta}_n^{(b)'} X_i)_+$. Let $\tilde{\beta}_n^{(b)}$ be the minimizer of $Q_n^{(b)}(\beta) \equiv \mathbb{P}_n^{(b)} (nL(X, Y, \beta) + \lambda_n \widehat{W}^{(b)} |\beta' X|)$. Let $\hat{u}$ and $\hat{l}$ be the $1 - \gamma/2$ and $\gamma/2$ percentiles of $\sqrt{n}(\mathbb{P}_n^{(b)} - \mathbb{P}_n) 1_{YX'\beta^{(b)} < 0}$. Then the $1 - \gamma$ centered percentile bootstrap (CPB) interval is given by $[\widehat{\tau}(\tilde{c}) - \hat{u}/\sqrt{n}, \widehat{\tau}(\tilde{c}) - \hat{l}/\sqrt{n}]$.

Theorem 3.3. *Assume the same setting as Theorem 3.2. Then confidence intervals for the test error $\tau(\tilde{c})$, based on either normal approximation or centered bootstrap percentiles, are asymptotically consistent.*

See the proof given in the Appendix.

4 Discussion

Inspired by the non-regular framework studied in Laber and Murphy, we proposed a family of adaptive classifiers. We discussed briefly some of the asymptotic properties of these classifiers, but many interesting questions remain open. Both theoretical investigation and simulations are needed in order to understand the small sample size behavior of the classifiers, their test error, and the test error confidence intervals. The best way to choose the adaptive weights is also an important open question. We are currently pursuing many of these avenues. We thank Laber and Murphy for their very interesting work and stimulation of important new research.

Acknowledgments

The second author was funded in part by grant CA142538 from the National Cancer Institute.

A Proofs

Proof of Theorem 3.2. We first assume that condition (NR) holds. Write $\beta = \beta^* + \frac{u}{\sqrt{n}}$ and Define

$$\Gamma_n(u) = Q_n\left(\beta^* + \frac{u}{\sqrt{n}}\right) - Q_n(\beta^*).$$

Let $\tilde{\mathbf{u}}_n = \operatorname{argmin}_{\mathbf{u}} \Gamma_n(\mathbf{u})$; then $\tilde{\mathbf{u}}_n = \sqrt{n}(\tilde{\beta}_n - \beta^*)$.

Fix $\mathbf{u}$ and let $G_n = \sum_{i=1}^n g(x_i, y_i, \beta^*)$ where g is the subgradient of L . Write

$$\Gamma_n(\mathbf{u}) = \sum_{i=1}^n \left(L \left(x_i, y_i, \beta^* + \frac{\mathbf{u}}{\sqrt{n}} \right) - L(x_i, y_i, \beta^*) \right) + \frac{\lambda_n}{n} \sum_{i=1}^n \widehat{w}_{ni} \left(\left| \left(\beta^* + \frac{\mathbf{u}}{\sqrt{n}} \right)' x_i \right| - |\beta^{*'} x_i| \right) \equiv A_n(\mathbf{u}) + B_n(\mathbf{u}). \quad (9)$$

It follows from Niemiro (1992, Eq. 4.12) that

$$A_n(\mathbf{u}) = \frac{\mathbf{u}'}{\sqrt{n}} G_n + nQ \left(\beta^* + \frac{\mathbf{u}}{\sqrt{n}} \right) + o_p(1) \rightarrow_d N \left(\frac{1}{2} \mathbf{u}' H \mathbf{u}, \Omega \right), \quad (10)$$

where Q , H , and Ω are defined in Laber and Murphy (2011, Section 3).

Consider now $B_n(\mathbf{u})$. Write

$$B_n(\mathbf{u}) = \frac{\lambda_n}{n} \sum_{i \in M_n} \widehat{w}_{ni} \left(\left| \left(\beta^* + \frac{\mathbf{u}}{\sqrt{n}} \right)' x_i \right| - |\beta^{*'} x_i| \right) + \frac{\lambda_n}{n} \sum_{i \in M_n^c} \frac{\widehat{w}_{ni}}{\sqrt{n}} |\mathbf{u}' x_i| \equiv B_{n1}(\mathbf{u}) + B_{n2}(\mathbf{u}). \quad (11)$$

Note that since $\lambda_n = O_p(1)$, it follows from (2) that $\max_{i \in M_n} \frac{\lambda_n}{\sqrt{n}} \widehat{w}_{ni} = o_p(1)$. Define

$T_{ni} = \sqrt{n} \left(\left| \left(\beta^* + \frac{\mathbf{u}}{\sqrt{n}} \right)' x_i \right| - |\beta^{*'} x_i| \right)$. Using the fact that for every two scalars a and b , $||a| - |b|| \leq |a - b|$, together with the Cauchy-Schwarz inequality, we obtain that $|T_{ni}| \leq \|\mathbf{u}\| \|x_i\|$. Then

$$|B_{n1}(\mathbf{u})| = \left| \frac{1}{n} \sum_{i \in M_n} \frac{\lambda_n}{\sqrt{n}} \widehat{w}_{ni} T_{ni} \right| \leq \left(\max_{i \in M_n} \frac{\lambda_n}{\sqrt{n}} \widehat{w}_{ni} \right) \frac{1}{n} \sum_{i \in M_n} |T_{ni}| \leq \left(\max_{i \in M_n} \frac{\lambda_n}{\sqrt{n}} \widehat{w}_{ni} \right) \|\mathbf{u}\| M = o_p(1), \quad (12)$$

where we used Assumption (B1) in the last inequality.

Consider now B_{n2} , the second expression in (11). Note that if $\mathbf{u} \in V$, then $B_{n2} \equiv 0$. Assume that $\mathbf{u} \notin V$. Note that for all $i \in M_n^c$

$$\frac{\widehat{w}_{ni}}{\sqrt{n}} = n^\alpha (1 - n^\alpha |(\widehat{\beta}_n - \beta^*)' x_i|) \geq n^\alpha (1 - n^\alpha \|\widehat{\beta}_n - \beta^*\| M).$$

Write

$$B_{n2}(\mathbf{u}) = \frac{\lambda_n}{n} \sum_{i \in M_n^c} \widehat{w}_{ni} |\mathbf{u}' x_i| = \lambda_n \frac{\# \{M_n^c\}}{n} \cdot \frac{1}{\# \{M_n^c\}} \sum_{i \in M_n^c} \frac{\widehat{w}_{ni}}{\sqrt{n}} |\mathbf{u}' x_i| \geq \lambda_n \frac{\# \{M_n^c\}}{n} \cdot n^\alpha (1 - n^\alpha \|\widehat{\beta}_n - \beta^*\| M) \cdot \frac{1}{\# \{M_n^c\}} \sum_{i \in M_n^c} |\mathbf{u}' x_i|. \quad (13)$$

Note that λ_n is bounded away from zero, and $\# \{M_n^c\}/n \rightarrow q$ by condition (NR). By the proof of Lemma 2.1, $n^\alpha (1 - n^\alpha \|\widehat{\beta}_n - \beta^*\| M) \rightarrow_p \infty$. By the law of large numbers and Assumption

(B3), $\frac{1}{\# \{M_n^c\}} \sum_{i \in M_n^c} |\mathbf{u}' x_i| \rightarrow c(\mathbf{u}) > 0$ where $c = c(\mathbf{u})$ is some positive constant. Using Slutsky's Theorem (Kosorok 2008, Theorem 7.15) we obtain that when $\mathbf{u} \notin V$, $B_{n2} \rightarrow_p \infty$.

Summarizing,

$$B_n(\mathbf{u}) \rightarrow \begin{cases} 0 & \mathbf{u} \in V \\ \infty & \mathbf{u} \notin V \end{cases}.$$

Thus,

$$\Gamma_n(\mathbf{u}) \rightarrow_d \Gamma(\mathbf{u}) \equiv \begin{cases} \frac{1}{2} \mathbf{u}' H \mathbf{u} + \mathbf{Z}' \mathbf{u} & \mathbf{u} \in V \\ \infty & \mathbf{u} \notin V, \end{cases}$$

where $\mathbf{Z} \sim N_p(\mathbf{0}, \Omega)$.

Note that in order to minimize $\Gamma(\mathbf{u})$, $\mathbf{u}$ must be in V . Equivalently $\mathbf{u} = \alpha \mathbf{v}$ for some α .

Thus, the the minimizer of $\frac{1}{2} \mathbf{u}' H \mathbf{u} + \mathbf{Z}' \mathbf{u}$ is the vector $\alpha \mathbf{v}$ for α that minimizes $\frac{\alpha^2}{2} \mathbf{v}' H \mathbf{v} + \alpha \mathbf{Z}' \mathbf{v}$. The minimizer is given by $-(\mathbf{v}' H \mathbf{v})^{-1} P_V \mathbf{Z}$.

We would like to show that $\tilde{\mathbf{u}}_n \rightarrow_d \tilde{\mathbf{u}}$, where $\tilde{\mathbf{u}}$ is the minimizer of $\Gamma(\mathbf{u})$. Note that $\Gamma_n(\mathbf{u})$ for all $n \geq 1$ and $\Gamma(\mathbf{u})$ are stochastic processes indexed by $\mathbb{R}^p$. The sample paths of Γ are lower semicontinuous and possess a unique minimum. We would like to show that $\{\tilde{\mathbf{u}}_n\}_n$ is uniformly tight. It is enough to show that for any given $\varepsilon > 0$, there exists a constant C_1 such that

$$P \left(\inf_{\|\mathbf{u}\|=C_1} Q_n(\beta^* + \mathbf{u} / \sqrt{n}) > Q_n(\beta^*) \right) \geq 1 - \varepsilon \quad (14)$$

for all n large enough. The result then follows from the convexity of Q_n . It follows from Niemiro (1992, Eq. 4.12) that

$$Q_n \left(\beta^* + \frac{\mathbf{u}}{\sqrt{n}} \right) - Q_n(\beta^*) \geq \frac{\mathbf{u}'}{\sqrt{n}} \sum_{i=1}^n g(x_i, y_i, \beta^*) + \left(\frac{1}{2} \mathbf{u}' H \mathbf{u} + \mathbf{o}_p(1) \right) + \frac{\lambda_n}{n} \sum_{i \in M_n} \widehat{w}_{ni} \left(\left| \left(\beta^* + \frac{\mathbf{u}}{\sqrt{n}} \right)' x_i \right| - |\beta^{*'} x_i| \right).$$

The first term is $O_p(1)$. By Assumption (A5), H is positive definite and hence for C_1 large enough the second term dominates the first term. By (12), the third term is also dominated by the second term and thus (14) holds. Thus we obtain that $\{\tilde{\mathbf{u}}_n\}_n$ is uniformly tight. Hence, all the conditions of the Argmax Theorem (Kosorok 2008, Theorem 14.1) hold, and consequently, $\tilde{\mathbf{u}}_n \rightarrow_d \tilde{\mathbf{u}}$. **Summarizing, we have $\tilde{\mathbf{u}}_n \rightarrow_d -(\mathbf{v}' H \mathbf{v})^{-1} P_V \mathbf{Z}$. In other words, $\sqrt{n}(\tilde{\beta}_n - \beta^*) \rightarrow_d N(\mathbf{0}, (\mathbf{v}' H \mathbf{v})^{-2} P_V \Omega P_V)$, which concludes the proof of (7).**

We would like to show that the oracle property holds, i.e., that $\tilde{\beta}_n \in V$ with probability that tends toward 1. It is sufficient to show that with probability tending to 1 as $n \rightarrow \infty$, for any β satisfying $\beta - \beta^* = O_p(n^{-1/2})$ and $\beta \notin V$, $Q_n(\beta) > Q_n(P_V \beta)$. By the definition of the subgradient, (see, for example, Niemiro 1992, page 1519), we have

$$(\beta - P_V \beta)' \sum_{i=1}^n g(x_i, y_i, P\beta) \leq n \mathbb{P}_n L(X, Y, \beta) - n \mathbb{P}_n L(X, Y, P\beta). \quad (15)$$

Write $\beta = \beta^* + \frac{\mathbf{u}_V}{\sqrt{n}} + \frac{\mathbf{u}_{V^\perp}}{\sqrt{n}}$, and note that $\beta - P_V \beta = \frac{\mathbf{u}_{V^\perp}}{\sqrt{n}}$. Note also that by (15)

$$\begin{aligned} \frac{Q_n(\beta) - Q_n(P_V \beta)}{\|\mathbf{u}_{V^\perp}\|} &\geq \|\mathbf{u}_{V^\perp}\|^{-1} \left((\beta - P_V \beta)' \sum_{i=1}^n g(x_i, y_i, P_V \beta) + \frac{\lambda_n}{n} \sum_{i=1}^n \widehat{w}_{ni} (|\beta' x_i| - |P_V \beta' x_i|) \right) \\ &= \frac{1}{\sqrt{n}} \frac{\mathbf{u}_{V^\perp}}{\|\mathbf{u}_{V^\perp}\|} \sum_{i=1}^n g \left(x_i, y_i, \beta^* + \frac{\mathbf{u}_V}{\sqrt{n}} \right) + \frac{\lambda_n}{n} \frac{1}{\|\mathbf{u}_{V^\perp}\|} \sum_{i \in M_n} \widehat{w}_{ni} \left(|\beta' x_i| - \left| \left(\beta^* + \frac{\mathbf{u}_V}{\sqrt{n}} \right)' x_i \right| \right) + \frac{\lambda_n}{n} \sum_{i \in M_n^c} \widehat{w}_{ni} \frac{|\mathbf{u}_{V^\perp}' x_i|}{\|\mathbf{u}_{V^\perp}\|} \\ &\equiv \tilde{A}_n + \tilde{B}_{n1} + \tilde{B}_{n2}. \end{aligned}$$

By Assumptions (A4)–(A5), $\tilde{A}_n = O_p(1)$ uniformly in $\mathbf{u}$. Using similar arguments to those used to prove (12), $\tilde{B}_{n1} = o_p(1)$ uniformly in $\mathbf{u}$. Finally, similarly to the arguments in (13), $\tilde{B}_{n2} \rightarrow_p \infty$ uniformly in $\mathbf{u}$. Hence $Q_n(\beta) > Q_n(P_V \beta)$ with probability that tends toward one uniformly in $\mathbf{u}$. This concludes the proof of (8).

Until now we assumed that condition (NR) holds. When condition (NR) does not hold, we obtain that $\Gamma_n(\mathbf{u}) = A_n(\mathbf{u}) + B_{n1}(\mathbf{u})$, where A_n is defined in (10) and B_{n1} is defined in (11). It

follows from the same arguments as above that $A_n(\mathbf{u}) \rightarrow_d N(\frac{1}{2} \mathbf{u}' H \mathbf{u}, \Omega)$ and $B_{n1}(\mathbf{u}) \rightarrow_p 0$. Hence,

$$\Gamma_n(\mathbf{u}) \rightarrow_d \frac{1}{2} \mathbf{u}' H \mathbf{u} + Z' \mathbf{u},$$

where $Z \sim N_p(\mathbf{0}, \Omega)$. Note that $\tilde{\mathbf{u}}$, the minimizer of $\frac{1}{2} \mathbf{u}' H \mathbf{u} + Z' \mathbf{u}$, is $-H^{-1} Z$ and has distribution $N(\mathbf{0}, H^{-1} \Omega H^{-1})$. It can be shown that the Argmax theorem conditions hold and thus we obtain that $\sqrt{n}(\tilde{\beta}_n - \beta^*) = \tilde{\mathbf{u}}_n \rightarrow_d \tilde{\mathbf{u}}$. This proves (7) for the case that (NR) does not hold.

Proof of Theorem 3.3. We first prove that the normal approximation confidence interval is consistent. Recall that (see Laber and Murphy 2011, Eq. 6),

$$\mathbb{G}_n 1_{YX' \tilde{\beta}_n < 0} = \mathbb{G}_n 1_{X' \beta^* = 0} 1_{YX' \tilde{\beta}_n < 0} + \mathbb{G}_n 1_{X' \beta^* \neq 0} 1_{YX' \tilde{\beta}_n < 0}.$$

When condition (NR) does not hold, $\mathbb{G}_n 1_{X' \beta^* = 0} 1_{YX' \tilde{\beta}_n < 0} \equiv 0$ almost surely. When condition (NR) holds, it follows from (8), that with probability tending toward 1, $P 1_{X' \beta^* = 0} 1_{YX' \tilde{\beta}_n < 0} = P_n 1_{X' \beta^* = 0} 1_{YX' \tilde{\beta}_n < 0} = 0$. Since $\tilde{\beta}_n \rightarrow \beta^*$, it follows from the proof of Theorem 3.1 of Laber and Murphy (2011) that

$$\mathbb{G}_n 1_{X' \beta^* \neq 0} 1_{YX' \tilde{\beta}_n < 0} \rightarrow_d \mathbb{B}(\beta^*),$$

where $\mathbb{B}(\beta^*)$ is normal with zero expectation and σ_1^2 variance. Since $\widehat{\sigma}_1^2$ is consistent for σ_1^2 , the normal approximation confidence interval is consistent.

We now move to prove consistency of the CPB. This proof has the same structure of the proof of Theorem 3.2. Hence, in the following, we mainly focus on the details that are different in the bootstrap proof.

It can be shown that Lemma 2.1 holds for the bootstrap weights. In other words, $\max_{i \in M_n^{(b)}} \widehat{w}_{ni}^{(b)} / \sqrt{n} = o_p(1)$, and for all $i \in M_n^{(b)c}$, $\widehat{w}_{ni}^{(b)} / \sqrt{n} \rightarrow_p \infty$. This follows by the same arguments used in the lemma's proof, together with the fact that $\sqrt{n}(\widehat{\beta}_n^{(b)} - \beta^*) = O_p(n^{-1/2})$.

Assume first that condition (NR) holds. Fix $\mathbf{u}$ and write $\Gamma_n^{(b)}(\mathbf{u}) = Q_n^{(b)}(\beta^* + \frac{\mathbf{u}}{\sqrt{n}}) - Q_n^{(b)}(\beta^*)$. Write $\Gamma_n^{(b)}(\mathbf{u}) \equiv A_n^{(b)}(\mathbf{u}) + B_{n1}^{(b)}(\mathbf{u}) + B_{n2}^{(b)}(\mathbf{u})$ where $A_n^{(b)}, B_{n1}^{(b)}, B_{n2}^{(b)}$, similarly to (10) and (11). Note that

$$\begin{aligned} A_n^{(b)}(\mathbf{u}) &= \frac{\mathbf{u}'}{\sqrt{n}} G_n^{(b)} + nQ \left(\beta^* + \frac{\mathbf{u}}{\sqrt{n}} \right) + o_p(1) \\ &= \frac{\mathbf{u}'}{\sqrt{n}} (G_n^{(b)} - G_n) + \frac{\mathbf{u}'}{\sqrt{n}} G_n + nQ \left(\beta^* + \frac{\mathbf{u}}{\sqrt{n}} \right) + o_p(1) \\ &= \mathbf{u}'(Z_{n1} + Z_{n2}) + nQ \left(\beta^* + \frac{\mathbf{u}}{\sqrt{n}} \right) + o_p(1) \rightarrow_d \frac{1}{2} \mathbf{u}' H \mathbf{u} + \mathbf{u}'(Z_1 + Z_2), \end{aligned}$$

where $Z_{n1} = G_n^{(b)} - G_n$, $Z_{n2} = G_n$, $Z_i \sim N(\mathbf{0}, \Omega)$ for $i = 1, 2$, and $Z_1 \perp Z_2$. Using the bootstrap version of Lemma 2.1, together with (12) and (13), we obtain that $B_{n1}^{(b)}(\mathbf{u}) = o_p(1)$, and $B_{n2}^{(b)}(\mathbf{u})$ converges in probability to zero for all $\mathbf{u} \in V$, and to infinity, otherwise. Thus, $\Gamma_n(\mathbf{u}) \rightarrow_d$

$\Gamma^{(b)}(\mathbf{u})$ where $\Gamma^{(b)}(\mathbf{u}) = \frac{1}{2} \mathbf{u}' H \mathbf{u} + \mathbf{u}'(Z_1 + Z_2)$ for $\mathbf{u} \in V$ and $\Gamma^{(b)}(\mathbf{u}) = \infty$ otherwise. The minimizer $\tilde{\mathbf{u}}^{(b)}$ of $\Gamma^{(b)}(\mathbf{u})$ is given by $-P_V H^{-1} P_V (Z_1 + Z_2)$. By the Argmax Theorem,

$\tilde{\mathbf{u}}_n^{(b)} \rightarrow_d \tilde{\mathbf{u}}^{(b)}$ or, similarly, $\sqrt{n}(\tilde{\beta}_n^{(b)} - \beta^*) = -P_V H^{-1} P_V (Z_{n1} + Z_{n2}) + o_p(1) \rightarrow -P_V H^{-1} P_V (Z_1 + Z_2)$. By the proof of Theorem 3.2, we also have that

$\sqrt{n}(\tilde{\beta}_n - \beta^*) = -P_V H^{-1} P_V Z_{n2} + o_p(1)$, and hence $\sqrt{n}(\tilde{\beta}_n^{(b)} - \tilde{\beta}_n) = -P_V H^{-1} P_V Z_{n1} + o_p(1) \rightarrow_d -P_V H^{-1} P_V Z_1$ unconditionally. When condition (NR) does not hold, we obtain, using similar arguments, that $\sqrt{n}(\tilde{\beta}_n^{(b)} - \tilde{\beta}_n) \rightarrow_d H^{-1} Z_1$ unconditionally. Standard arguments can be used to show that the unconditional convergence implies conditional convergence, and thus the bootstrap is consistent.

References

- Fan J, Li R. Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association*. 2001; 96(456):1348–1360.
- Kosorok, MR. *Introduction to Empirical Processes and Semiparametric Inference*. Springer; 2008.
- Laber EB, Murphy SA. Adaptive Confidence Intervals for the Test Error in Classification. *Journal of the American Statistical Association*. 2011 To appear.
- Niemiro W. Asymptotics for M-Estimators Defined by Convex Minimization. *The Annals of Statistics*. 1992; 20(3):1514–1533.
- Zou H. The Adaptive Lasso and its Oracle Properties. *Journal of the American Statistical Association*. 2006; 101(476):1418–1429.