

Supplementary Materials for “Least Ambiguous Set-Valued Classifiers with Bounded Error Levels”

A Proofs

We present the proofs of the results that are not clear in the main article.

Lemma 5. Proof. Let \mathbf{H} and \mathbf{H}' be such that $\mathbb{P}\{Y \in \mathbf{H}(X)|Y = y\} = \mathbb{P}\{Y \in \mathbf{H}'(X)|Y = y\} = 1 - \alpha_y$, and $\mathbb{P}\{y \in \mathbf{H}'(X)|Y \neq y\} \geq \mathbb{P}\{y \in \mathbf{H}(X)|Y \neq y\}$, for all y . Multiplying this expression by $\mathbb{P}(Y \neq y)$ we obtain $\mathbb{P}\{y \in \mathbf{H}'(X), Y \neq y\} \geq \mathbb{P}\{y \in \mathbf{H}(X), Y \neq y\}$, which can be rewritten as

$$\sum_{l \neq y} \mathbb{P}\{y \in \mathbf{H}'(X)|Y = l\}\pi_l \geq \sum_{l \neq y} \mathbb{P}\{y \in \mathbf{H}(X)|Y = l\}\pi_l, \quad (9)$$

which holds for all y . On the other hand we have

$$\sum_y \mathbb{P}\{Y \in \mathbf{H}'(X)|Y = y\}\pi_y = \sum_y \mathbb{P}\{Y \in \mathbf{H}(X)|Y = y\}\pi_y = \sum_y (1 - \alpha_y)\pi_y.$$

Adding Expression (9) over all y and combining with the last expression leads to

$$\sum_y \sum_l \mathbb{P}\{y \in \mathbf{H}'(X)|Y = l\}\pi_l \geq \sum_y \sum_l \mathbb{P}\{y \in \mathbf{H}(X)|Y = l\}\pi_l,$$

which by the law of total probability and Remark 4 is equivalent to $\mathbb{E}\{|\mathbf{H}'(X)|\} \geq \mathbb{E}\{|\mathbf{H}(X)|\}$. \square

Theorem 6. Proof. First, notice that $\text{logit}\{p(y|x)\} = \log\{p(x|y)/p(x|y^c)\} + \text{logit}(\pi_y)$, where $p(x|y^c) \equiv \sum_{j \neq y} p(x|Y = j)\pi_j / \sum_{j \neq y} \pi_j$. Given that the log and logit functions are monotonically increasing, this expression implies that the decision regions C_y can alternatively be based on level sets of the likelihood ratios

$\Lambda_y(x) = p(x|y)/p(x|y^c)$, that is $C_y = \{x : \Lambda_y(x) \geq \ell_y\}$ with ℓ_y chosen so that $\mathbb{P}(C_y|Y = y) = 1 - \alpha_y$. The region C_y^c therefore corresponds to the Neyman-Pearson rejection region for testing the null hypothesis $H_0 : Y = y$ versus $H_1 : Y \neq y$. By the Neyman-Pearson lemma we have that the classifier \mathbf{H} induced by the sets C_y maximizes the probabilities $\mathbb{P}\{y \notin \mathbf{H}(X)|Y \neq y\}$, or equivalently $\mathbb{P}\{y \in \mathbf{H}(X)|Y \neq y\}$ is minimized. Finally, by Lemma 5 we have that this decision rule \mathbf{H} also minimizes the ambiguity. \square

Theorem 8. Proof. Firstly, since $\tilde{\mathbb{A}}$ is the optimal value of problem (3), $\tilde{\mathbb{A}} \leq \mathbb{A}(\mathbf{H}^\dagger)$. Now, $\mathbb{A}(\mathbf{H}^\dagger) = \mathbb{E}[I\{X \in \mathcal{N}(\mathbf{H}^*)\}|\mathbf{H}^\dagger(X)|] + \mathbb{E}[I\{X \notin \mathcal{N}(\mathbf{H}^*)\}|\mathbf{H}^\dagger(X)|] = \mathbb{P}\{\mathcal{N}(\mathbf{H}^*)\} + \mathbb{A}(\mathbf{H}^*)$, and the result follows from $\mathbb{A}(\mathbf{H}^*) \leq \tilde{\mathbb{A}}$ given that (2) is a relaxation of (3). \square

Theorem 14. Proof. The first part is essentially the same as in Lei (2014). We prove the second part. Let \hat{G}_y be the empirical distribution of $p(y|X_{y,1}), \dots, p(y|X_{y,n_y})$ where $X_{y,1}, \dots, X_{y,n_y}$ are sample points in class y . Let $\hat{\mathbb{P}}_y(\cdot)$ be the probability measure corresponding to \hat{G}_y . Define $L_y(t) = \{x : p(y|x) \leq t\}$, $\hat{L}_y(t) = \{x : \hat{p}(y|x) \leq t\}$.

We focus on the event

$$E = \left\{ \sup_{y,x} |\hat{p}(y|x) - p(y|x)| \leq \epsilon_n, \sup_{y,t} |\hat{G}_y(t) - G_y(t)| \leq c\sqrt{\frac{\log n}{n}}, \right. \\ \left. \sup_y |\hat{\pi}_y - \pi_y| \leq c\sqrt{\frac{\log n}{n}} \right\},$$

which has probability at least $1 - K\delta_n - n^{-1}$ if c is chosen large enough and K grows slowly with n . Here the first inequality in E is given by our assumption in (4) and the other two follow from standard empirical process theory.

Recall that for total coverage we use the same threshold for all classes. Let $t^* = G^{-1}(\alpha)$ be the ideal cut-off value for $p(y|x)$. If $t \leq t^* - \epsilon_n - \{(K+1)cc_1^{-1}\sqrt{\log n/n}\}^{1/\gamma}$, then

we have

$$\begin{aligned}\widehat{\mathbb{P}}_y\{\widehat{L}_y(t)\} &\leq \widehat{\mathbb{P}}_y\{L(t + \epsilon_n)\} = \widehat{G}_y(t + \epsilon_n) \leq G_y(t + \epsilon_n) + c\sqrt{\frac{\log n}{n}} \\ &\leq G_y\left[t^* - \{(K+1)cc_1^{-1}\sqrt{\log n/n}\}^{1/\gamma}\right] + c\sqrt{\frac{\log n}{n}} \leq G_y(t^*) - cK\sqrt{\frac{\log n}{n}}.\end{aligned}$$

Therefore,

$$\widehat{t} > t^* - \epsilon_n - \left\{(K+1)cc_1^{-1}\sqrt{\log n/n}\right\}^{1/\gamma}, \quad (10)$$

because otherwise we have

$$\begin{aligned}\sum_{y=1}^K \widehat{\pi}_y \widehat{\mathbb{P}}_y\{\widehat{L}_y(\widehat{t})\} &\leq \sum_{y=1}^K \widehat{\pi}_y \{G_y(t^*) - cK\sqrt{\log n/n}\} \\ &\leq \alpha + \sum_{y=1}^K |\widehat{\pi}_y - \pi_y| G_y(t^*) - cK\sqrt{\log n/n} < \alpha.\end{aligned}$$

Similarly we can obtain

$$\widehat{t} \leq t^* + \epsilon_n + \left\{(K+1)cc_1^{-1}\sqrt{\log n/n}\right\}^{1/\gamma}, \quad (11)$$

and combining (10) and (11) we have $|\widehat{t} - t^*| \leq \epsilon_n + \left\{(K+1)cc_1^{-1}\sqrt{\log n/n}\right\}^{1/\gamma}$. (It is worth noting that a rigorous argument of this would require $\widehat{p}(y|x)$ to have distinct values at the sample points X_1, \dots, X_n . This is a minor issue because one can always add very small random perturbations such as $\widehat{p}(y|X) + \xi$ with $\xi \sim \text{Unif}(-n^{-2}, n^{-2})$.)

Then

$$\begin{aligned}\mathbb{P}_y\left(\widehat{C}_y \setminus C_y^*\right) &= \mathbb{P}_y\left\{\widehat{p}(y|X) \geq \widehat{t}, p(y|X) < t^*\right\} \\ &\leq \mathbb{P}_y\left[t^* - 2\epsilon_n - \left\{(K+1)cc_1^{-1}\sqrt{\log n/n}\right\}^{1/\gamma} \leq p(y|X) < t^*\right] \\ &\leq c'\left(\epsilon_n^\gamma + K\sqrt{\log n/n}\right),\end{aligned}$$

for some constant c' depending on c, c_1, γ . Similarly we can obtain $\mathbb{P}_y(C_y^* \setminus \hat{C}_y) \leq c'(\epsilon_n^\gamma + K\sqrt{\log n/n})$, and hence $\mathbb{P}_y(\hat{C}_y \triangle C_y^*) \leq c'(\epsilon_n^\gamma + K\sqrt{\log n/n})$. Summing over y we have

$$\mathbb{P}(\hat{\mathbf{H}} \triangle \mathbf{H}^*) = \sum_{y=1}^K \pi_y \mathbb{P}_y(\hat{C}_y \triangle C_y^*) \leq c'(\epsilon_n^\gamma + K\sqrt{\log n/n}).$$

□

B Simulation Studies

B.1 Univariate Scenarios

We start with a simple setting to illustrate the fundamental differences between classifiers with reject option (CWRs) and LABEL classifiers. In this comparison we simulate samples of size $n = 4000$, drawing Y from $\{1, 2, 3\}$ with probabilities that change in three simulation scenarios, summarized in Table 2. We take X to be univariate with distributions $(X|Y = y)$ being normal with means $-2, 0$, and 2 , and variances equal to 1. We estimate $p(y|x)$ using $\hat{p}(y|x) = \hat{p}_y(x)\hat{\pi}_y / \sum_l \hat{p}_l(x)\hat{\pi}_l$, where $\hat{\pi}_y = \sum_i I(Y_i = y)/n$ and $\hat{p}_y(x)$ being a Gaussian density. We use total coverage of 0.95 for the plug-in LABEL and CWR classifiers. For each scenario we repeat the simulation 1000 times, and in Table 2 we report the average ambiguity and class coverages across simulations.

From Table 2 we can see that, across the three scenarios that we considered, the LABEL classifier has smaller average ambiguity than the CWR. In fact, the ambiguity of the LABEL classifier was smaller in all 1000 simulation replicates within each simulation scenario, not only on average. This is rather natural, since in our construction of LABEL classifiers assigning the output $\{1, 2, 3\}$ to a sample point is penalized

Class Probs.	CWR		LABEL	
	Ambiguity	Class Coverage	Ambiguity	Class Coverage
(.45, .10, .45)	1.28	(1.00, 0.53, 1.00)	1.21	(0.98, 0.64, 0.98)
(.33, .33, .34)	1.94	(0.98, 0.89, 0.98)	1.51	(0.96, 0.93, 0.96)
(.60, .30, .10)	1.78	(0.99, 0.89, 0.92)	1.41	(0.98, 0.91, 0.88)

Table 2: Comparing LABEL classifiers with total coverage control to CWRs in a three-class problem. Simulation details are given in the main text. Quantities reported here are averages over 1000 simulation replicates.

more than an output containing only two labels. CWRs effectively assign the output $\{1, 2, 3\}$ to all ambiguous sample points, and therefore do not specify which labels are plausible for a given instance. In the scenarios explored in our simulation study, classes 1 and 3 are relatively well separated, and therefore the LABEL classifier assigns the outcomes $\{1, 2\}$ and $\{2, 3\}$ to the sample points in the overlap of classes 1 and 2, and 2 and 3, respectively. To such cases, the CWR assigns $\{1, 2, 3\}$, which is less informative and leads to larger ambiguity. LABEL classifiers are therefore more informative in reporting ambiguous cases as they indicate the set of plausible labels for each instance.

We also conclude from Table 2 that controlling total coverage can lead to very uneven class coverage with both LABEL classifiers and CWRs. Notwithstanding, our framework can also be used to control class-specific coverage, something that cannot be done using CWRs.

	Ambiguity	Class Coverage
CWR	3.64	(0.87, 0.98, 0.93, 0.92, 0.97)
LABEL	1.87	(0.90, 0.99, 0.97, 0.69, 0.53)

Table 3: Comparing LABEL classifiers and CWRs under total coverage control in a multivariate problem. Simulation details are given in the main text. Quantities reported here are averages over 1000 simulation replicates.

B.2 Data-Based Multivariate Scenario

To explore the performance of LABEL classifiers in comparison with CWRs under more complex scenarios, we use the Abalone data analyzed in Section 5.4 to create a synthetic population from which we simulate. The goal here is to construct a synthetic population based on real data, not necessarily study the repeated sample performance of the analysis in Section 5.4. We start by creating $K = 5$ classes using the age variable (unlike in Section 5.4, where we took $K = 3$ to ease the presentation). We then use the seven numeric features in the Abalone data to obtain a mean vector and a covariance matrix within each class. The population is then defined as a five-component mixture of seven-dimensional normal distributions, with means, covariance matrices, and population proportions obtained from the abalone data.

Each of the 1000 replicates in this simulation study consists of $n = 4000$ draws from the aforementioned mixture. The estimator of $p(y|x)$ is a simple multinomial logistic regression. We controlled the total coverage of the classifiers at 0.95. For CWRs this means that we chose the cost of the reject option ρ so that its total coverage is greater or equal to 0.95.

In Table 3 we compare the ambiguity and class coverages that we obtain on average

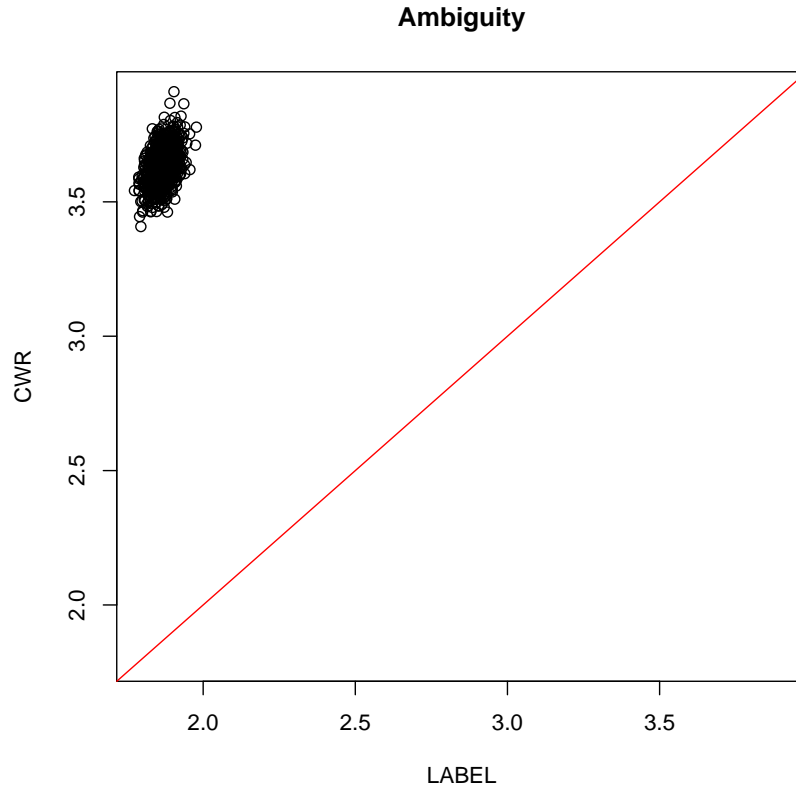


Figure 7: Comparison of ambiguity obtained from LABEL and CWR across 1000 simulation replicates in multivariate scenario.

from LABEL and CWR. Similarly as with the other examples and simulation scenarios, LABEL classifiers provide smaller ambiguity values, meaning that these are more informative than CWRs as the latter provide larger outputs. Indeed, in Figure 7 we show that LABEL ambiguity is smaller than CWR ambiguity in all simulation replicates.

In Table 3 we can also see that LABEL classifiers tend to give very imbalanced class coverages. In this simulation scenario this occurs because three of the classes have

very small class probabilities (classes 1, 4 and 5), and in addition two of these are not very well separated from a third one (classes 4 and 5 are very close to class 3). This phenomenon was illustrated in Example 3. Regarding CWRs, their larger outputs lead in this case to higher and more balanced class coverages given that instances receive all labels when they get assigned the reject option. Nevertheless, balance of class coverage cannot be guaranteed with CWRs, whereas LABEL classifiers can indeed be used under class-specific coverage control.