6. Appendix

6.1. Proof of Claim 1

Proof The expected conditional risk can be solved optimally by a dynamic program, where a DP recursion is,

\[
J_K(x^K, S^K) = \min_{f_K} E_y \left[ R_k(y, x^K, f^K, S^K) \right]
\]

\[
J_k(x^k, S^k) = \min_{f_k} E_y \left[ R_k(y, x^k, f^k, S^k) \right] + E_{x^{k+1}, \ldots, x^K} \left[ J_{k+1}(x^{k+1}, S^{k+1}) \right]
\]

Consider \( k \)th stage minimization, \( f_k \) can take three possible values \(+1, -1, r\) and \( J_k(x^k, S^k) \) can be recast as an conditional expected risk minimization,

\[
J_k(x^k, S^k) = S^k \min_{f_k} \left\{ w_n P_y [y = -1 | x^k], w_p P_y [y = +1 | x^k], \delta^k + E_{x^{k+1}, \ldots, x^K} \left[ J_{k+1}(x^{k+1}, S^{k+1} = 1) \right] \right\}
\]

Now, define \( \tilde{\delta}(x^k) = \delta^k + E_{x^{k+1}, \ldots, x^K} \left[ J_{k+1}(x^{k+1}, S^{k+1} = 1) \right] \) and solve the conditional risk above for \( S^k = 1 \),

\[
f_k(x^k) = \begin{cases} 
  -1, & \text{if } P(y = 1 | x^k) < \frac{\tilde{\delta}(x^k)}{w_p} \\
  r, & \text{if } \frac{\tilde{\delta}(x^k)}{w_p} \leq P(y = 1 | x^k) \leq 1 - \frac{\tilde{\delta}(x^k)}{w_n} \\
  +1, & \text{if } P(y = 1 | x^k) > 1 - \frac{\tilde{\delta}(x^k)}{w_n}
\end{cases}
\]

which is exactly our claim.

6.2. Proof of Claim 2

Proof The conditional expected risk for a given \( x \), \( \tilde{\delta}(x) \) and error penalties \( w_n, w_p \) is,

\[
\min \left\{ w_n P_y [y = -1 | x], w_p P_y [y = +1 | x], \tilde{\delta}(x) \right\}
\]

The optimal bayesian classifier that minimizes this risk is,

\[
f(x) = \begin{cases} 
  -1, & \text{if } P(y = 1 | x) < \frac{\tilde{\delta}(x)}{w_p} \\
  r, & \text{if } \frac{\tilde{\delta}(x)}{w_p} \leq P(y = 1 | x) \leq 1 - \frac{\tilde{\delta}(x)}{w_n} \\
  +1, & \text{if } P(y = 1 | x) > 1 - \frac{\tilde{\delta}(x)}{w_n}
\end{cases}
\]
We need to show that $f$ can be decomposed as a pair of binary classifiers $f_n, f_p : \mathcal{X} \rightarrow \{+1, -1\}$. Consider the following decomposition,

$$f(x) = \begin{cases} f_p(x), & f_p(x) = f_n(x) \\ r, & f_p(x) \neq f_n(x) \end{cases}$$ \hfill (29)

The conditional expected risk with this decomposition,

$$\min \left\{ w_n \mathbb{P}_y[y = -1 \mid x], w_p \mathbb{P}_y[y = +1 \mid x], \tilde{\delta}(x) \right\}$$ \hfill (30)

Note that the expected risk is symmetric and $f_n$ and $f_p$ can be interchanged. However, consider the equations for $f_p$ and $f_n$ that follow from minimizing the risk. Here, we used the fact that $\mathbb{P}(y = -1 \mid x) = 1 - \mathbb{P}(y = 1 \mid x)$.

$$f_p(x) = \begin{cases} +1, & \mathbb{P}(y = 1 \mid x) > \frac{\tilde{\delta}(x)}{w_p} \\ -1, & \mathbb{P}(y = 1 \mid x) \leq \frac{\tilde{\delta}(x)}{w_p} \end{cases}$$ \hfill (31)

$$f_n(x) = \begin{cases} +1, & \mathbb{P}(y = 1 \mid x) > 1 - \frac{\tilde{\delta}(x)}{w_n} \\ -1, & \mathbb{P}(y = 1 \mid x) \leq 1 - \frac{\tilde{\delta}(x)}{w_n} \end{cases}$$ \hfill (32)

Note that we chose our convention such that $f_p$ is positively biased classifier and $f_n$ is negatively biased classifier.

And, by inspection, (29) is true, therefore our decomposition is the optimal bayesian classifier.

Also, note another interesting observation, $f_p$ and $f_n$ are solutions to the following biased classification problems,

$$f_p = \arg \min \left\{ \frac{1 - \tilde{\delta}(x)}{w_p} \mathbb{P}_y[y = -1 \mid x], \frac{\tilde{\delta}(x)}{w_p} \mathbb{P}_y[y = +1 \mid x] \right\}$$ \hfill (33)

$$f_n = \arg \min \left\{ \frac{\tilde{\delta}(x)}{w_n} \mathbb{P}_y[y = -1 \mid x], 1 - \frac{\tilde{\delta}(x)}{w_n} \mathbb{P}_y[y = +1 \mid x] \right\}$$ \hfill (34)

Here, we used a standard Bayesian solution to a conditional expected risk for binary classification with weights $k$ and $1 - k$,

$$f^* = \arg \min \left\{ (1 - k) \mathbb{P}_y[y = -1 \mid x], (k) \mathbb{P}_y[y = +1 \mid x] \right\}$$ \hfill (36)

$$\hfill (37)$$
\[ f^*(x) = \begin{cases} +1, & P(y = 1|x) > k \\ -1, & P(y = 1|x) \leq 1 - k \end{cases} \] (38)

### 6.3. Proof of Theorem 1

**Proof** Since the risk is a smooth function of \( q_n, q_p, q^2 \), our algorithm solves the following by coordinate descent minimization over \( q_n, q_p, q^2 \):

\[
\min_{q_n, q_p, q^2} \hat{R}(f_n, f_p, f^2) \tag{39}
\]

s.t. \( f_p = \sum_{h_j \in \mathcal{H}_1} q_p^2 h_j(x_i) \), \( f_n = \sum_{h_j \in \mathcal{H}_1} q_n^2 h_j(x_i) \) (40)

\( f^2 = \sum_{h_j \in \mathcal{H}_2} q_j^2 h_j(x_i) \) (41)

therefore we are guaranteed to converge to a local minimum. 

### 6.4. Theorem 2 (Generalization Error Bound)

Our approach employs margin maximizing algorithm. (Masnadi-Shirazi and Vasconcelos (2009)) So it is appropriate to prove an error margin generalization bound for a two stage system:

**Theorem 2** Let \( D \) be a distribution on \( \mathcal{X} \times \{+1, -1\} \), and let \( S \) be a sample of \( m \) examples chosen independently at random according to \( D \), and a rejected subsample of size \( m_r \), \( S_r = \{ x \in S | f_p(x) \neq f_n(x) \} \) Assume that the base-classifier spaces \( \mathcal{H}_1 \) and \( \mathcal{H}_2 \) are finite, and let \( \delta > 0 \). Then with probability at least \( 1 - \delta \) over the random choice of the training set \( S \), all boosted classifiers \( f_n, f_p, f^2 \) satisfy the following bound for all \( \theta_1 > 0 \) and \( \theta_2 > 0 \):

\[
P_D[yf_n(x) \leq 0, yf_p(x) \leq 0] + P_D[yf_2(x) \leq 0, f_n(x) \neq f_p(x)] \leq P_S[yf_n(x) \leq \theta_1, yf_p(x) \leq \theta_1] + P_{S_r}[yf_2(x) \leq \theta_2] \\
+ \mathcal{O}\left(\frac{1}{\sqrt{m}}\left(\frac{\log m \log |\mathcal{H}_1|}{\theta_1} + \log \frac{1}{\delta}\right)\right) + \mathcal{O}\left(\frac{1}{\sqrt{m_r}}\left(\frac{\log m_r \log |\mathcal{H}_2|}{\theta_2} + \log \frac{1}{\delta}\right)\right) \tag{42}\]

**Proof** This will closely follow the proof of Theorem 1 in Bartlett et al. (1998). We have to bound two terms: \( P_D[yf_n(x) \leq \theta_1, yf_p(x) \leq \theta_1] \) and \( P_D[yf_2(x) \leq \theta_2, yf_n(x) \neq yf_p(x)] \)

**First Term** Let us bound the first term. Define \( C_N \) to be the set of unweighted averages over \( N \) elements from \( \mathcal{H}_1 \),

\[
C_N = \{ f : x \rightarrow \frac{1}{N} \sum_{i=1}^{N} h_i(x) | h_i \in \mathcal{H}_1 \} \tag{43}
\]
Any weighed classifier $f = \sum_h q_h h(x)$ can be approximated by drawing an element from $C_N$ by choosing $h_1...h_N$ with prob. $q_h$. We can express our first term as a sum of probabilities of disjoint events.

$$P_D [yf_p(x) \leq 0, yf_n(x) \leq 0] = \left\{ \begin{array}{ll}
P_D [yf_p(x) \leq 0, yf_n(x) \leq 0, yg_p(x) \leq \frac{\theta_1}{2}, yg_n(x) \leq \frac{\theta_1}{2}] & (44) \\
+ P_D [yf_p(x) \leq 0, yf_n(x) \leq 0, yg_p(x) \leq \frac{\theta_1}{2}, yg_n(x) > \frac{\theta_1}{2}] & (45) \\
+ P_D [yf_p(x) \leq 0, yf_n(x) \leq 0, yg_p(x) > \frac{\theta_1}{2}, yg_n(x) \leq \frac{\theta_1}{2}] & (46) \\
+ P_D [yf_p(x) \leq 0, yf_n(x) \leq 0, yg_p(x) > \frac{\theta_1}{2}, yg_n(x) > \frac{\theta_1}{2}] & (47) \end{array} \right.$$ 

Further, we can write,

$$P_D [yf_p(x) \leq 0, yf_n(x) \leq 0] \leq P_D \left[yg_p(x) \leq \frac{\theta_1}{2}, yg_n(x) \leq \frac{\theta_1}{2}\right]$$

$$\leq \left[ yf_p(x) \leq 0, yf_n(x) \leq 0, yg_p(x) > \frac{\theta_1}{2}, yg_n(x) \leq \frac{\theta_1}{2} \right]$$

The inequality holds for any $g_p, g_n$. We take the expected value of the right hand side wrt to the distribution $C$

$$P_D [yf_p(x) \leq 0, yf_n(x) \leq 0] \leq \mathbb{E}_C \left[P_D \left[yg_p(x) \leq \frac{\theta_1}{2}, yg_n(x) \leq \frac{\theta_1}{2}\right]\right]$$

$$\leq \mathbb{E}_D \left[P_c, c_n \left[yg_p(x) > \frac{\theta_1}{2}, yg_n(x) > \frac{\theta_1}{2} \mid yf_p(x) \leq 0, yf_n(x) \leq 0\right]\right]$$

The last term inside the expectation is the probability that an average of $N$ Bernoulli random variables is larger than its expectation, we use a concentration result from Equation (4) in Theorem 1 of Bartlett et al. (1998).

$$P_{c, c_n} \left[yg_p(x) > \frac{\theta_1}{2}, yg_n(x) > \frac{\theta_1}{2} \mid yf_p(x) \leq 0, yf_n(x) \leq 0\right] \leq \exp \left(-\frac{N\theta_1^2}{8}\right)$$

To bound the first we use the result from Equation (5) in Theorem 1 of Bartlett et al. (1998). If we set $\epsilon_N = \sqrt{(1/2m)\log((N + 1)|\mathcal{H}_1|^2N)/\delta_N}$, with probability at least $1 - \delta_N$,

$$P_D, C \left[yg_p(x) \leq \frac{\theta_1}{2}, yg_n(x) \leq \frac{\theta_1}{2}\right] \leq P_{S, C} \left[yg_p(x) \leq \frac{\theta_1}{2}, yg_n(x) \leq \frac{\theta_1}{2}\right] + \epsilon_N$$

for any choice of $\theta$ and every distribution $C$. Here, $P_S$ is probability taken with respect to a randomly drawn sample of size $m$ from $D$. 

21
By setting \( \delta \) then we do not have strong generalization.

on the reject classifier at first stage. So if very few examples make it to the second stage

\[ m \] number of training examples at that stage. An interesting observation is that

the error over the training set and a term that is inversely proportional to the margin and the

stage boosted classifiers than we can bound the generalization error by the empirical margin

and second stage error on rejected fraction. It states that if we are given a first and second

The error is a sum of two terms: first stage error of data that is not rejected

Discussion:

Collecting the two terms produces the desired result.

Second Term

Here we will bound the second term, \( \mathbb{P}_D[yf_2(x) \leq \theta_2, yf_n(x) \neq yf_p(x)] \)

Define a new distribution:

\[
D_r = \begin{cases} 
  cD(x, y), & \text{if } f_p(x) \neq f_n(x) \\
  0, & \text{if } f_p(x) = f_n(x) 
\end{cases}
\]

Rewrite:

\[
\mathbb{P}_D[yf_2(x) \leq \theta_2, yf_n(x) \neq yf_p(x)] \leq \mathbb{P}_D[yf_2(x) \leq \theta_2 \mid yf_n(x) \neq yf_p(x)] 
= \mathbb{P}_{D_r}[yf_2(x) \leq \theta_2] 
\]

Note that \( \mathcal{S}_r \) is an iid sample from \( D_r \). Using Theorem 1 in Bartlett et al. (1998),

\[
\mathbb{P}_{D_r}[yf_2(x) \leq 0] \leq \mathbb{P}_{\mathcal{S}_r}[yf_2(x) \leq \theta_2] + \mathcal{O} \left( \frac{1}{\sqrt{m}} \left( \frac{\log m \log |\mathcal{H}|}{\theta_2} + \log \frac{1}{\delta} \right)^{\frac{1}{2}} \right) 
\]

Collecting the two terms produces the desired result.

**Discussion:** The error is a sum of two terms: first stage error of data that is not rejected and second stage error on rejected fraction. It states that if we are given a first and second stage boosted classifiers than we can bound the generalization error by the empirical margin error over the training set and a term that is inversely proportional to the margin and the number of training examples at that stage. An interesting observation is that \( m_r \) depends on the reject classifier at first stage. So if very few examples make it to the second stage then we do not have strong generalization.
6.5. Additional Experiments

In medical diagnosis and threat detection, the penalty of false positives and false negatives \((w_n, w_p)\) is not equal. The experiment in Fig. 10 demonstrates our global algorithms in the biased scenario. For each reject cost \(\delta\), we compute an ROC curve. We also compute a corresponding average reject rate for each value of delta. This reject rate is averaged over the values \((w_n, w_p)\). So the highest reject rate corresponds to the best performance but also to the highest acquisition cost incurred by the system.

Figure 10: Two Stage ROC using the global surrogate method. Each ROC curve corresponds to a different value of reject cost \(\delta\). The legend displays average reject rate for \(\delta\)'s. Note, the red ROC corresponds to the centralized system (100% reject rate). For both experiments, very good performance can be achieved by requesting only 50% of instances to be measured at the second stage.