A. Appendix A

A.1. Proof of Theorem 1

The proof is by contradiction. Fix a distribution $\mathbb{P}$ satisfying conditional independence, and let $\mathbf{x}$ denote a fixed set of instances. Denote $\mathbb{P}(Y=1|x_i)=\eta_i$ and the optimal classifier by $\mathbf{s}^* \in \{0,1\}^n$. Suppose there exist indices j,k such that $s_j^*=1, s_k^*=0$ and $\eta_j < \eta_k$. Let $\mathbf{s}' \in \{0,1\}^n$ be such that $s'_j=0$ and $s'_k=1$, but identical to $\mathbf{s}^*$ otherwise i.e. $s_i^*=s'_i \ \forall i \in [n]\backslash\{j,k\}$. Note that $\sum_{i=1}^n s_i^* = \sum_{i=1}^n s_i^{'}$. For convenience, define:

$$\mathcal{U}^L(\mathbf{s};\mathbb{P}) := \mathbf{E}_{\mathbf{Y}\sim\mathbb{P}(.|\mathbf{x})} L(\mathbf{s},\mathbf{Y}) \ .$$

By optimality of $\mathbf{s}^*$,

$$\mathcal{U}^L(\mathbf{s}^*;\mathbb{P}) - \mathcal{U}^L(\mathbf{s}';\mathbb{P}) \leq 0. \tag{6}$$

Consider the LHS, $\mathcal{U}^L(\mathbf{s}^*;\mathbb{P}) - \mathcal{U}^L(\mathbf{s}';\mathbb{P})$ is equal to:

$$\sum_{\mathbf{y}\in\{0,1\}^n} P(\mathbf{y}|\mathbf{x})[L(\mathbf{s}^*,\mathbf{y}) - L(\mathbf{s}',\mathbf{y})] =$$
$$\sum_{\mathbf{y}\in\{0,1\}^n:y_j\neq y_k} P(\mathbf{y}|\mathbf{x})[L(\mathbf{s}^*,\mathbf{y}) - L(\mathbf{s}',\mathbf{y})]$$
$$+ \sum_{\mathbf{y}\in\{0,1\}^n:y_j= y_k} P(\mathbf{y}|\mathbf{x}) \underbrace{[L(\mathbf{s}^*,\mathbf{y}) - L(\mathbf{s}',\mathbf{y})]}_{(*)}$$

Note that when $y_j = y_k$, $\sum_{i=1}^n s_i^* y_i = \sum_{i=1}^n s'_i y_i$, so $L(\mathbf{s}^*,\mathbf{y}) - L(\mathbf{s}',\mathbf{y}) = 0$. It follows that the term $(*)$ equals 0.

Next we apply the representation of Proposition 1 with $v(\mathbf{s}) = \frac{1}{n}\sum_{i=1}^n s_i$ and $p(\mathbf{y}) = \frac{1}{n}\sum_{i=1}^n y_i$. Let $z \in \{0,1\}^{n-2}$ denote the vector corresponding to $n-2$ indices $\{y_i,\ i \in [n]\setminus\{j,k\}\}$, then $\mathcal{U}^L(\mathbf{s}^*;\mathbb{P}) - \mathcal{U}^L(\mathbf{s}';\mathbb{P})$ is given by:

$$\sum_{\mathbf{y}\in\{0,1\}^n:y_j\neq y_k} \mathbb{P}(\mathbf{y}|\mathbf{x})[L(\mathbf{s}^*,\mathbf{y}) - L(\mathbf{s}',\mathbf{y})] =$$
$$\sum_{\mathbf{z}\in\{0,1\}^{n-2}} \mathbb{P}(\mathbf{z}, y_j=1, y_k=0|\mathbf{x})\big[\Phi(\widehat{\mathrm{TP}}(\mathbf{s}^*,\mathbf{y}), v(\mathbf{s}^*), p(\mathbf{y}))$$
$$- \Phi(\widehat{\mathrm{TP}}(\mathbf{s}',\mathbf{y}), v(\mathbf{s}'), p(\mathbf{y}))\big]$$
$$+ \mathbb{P}(\mathbf{z}, y_j=0, y_k=1|\mathbf{x})\big[\Phi(\widehat{\mathrm{TP}}(\mathbf{s}^*,\mathbf{y}), v(\mathbf{s}^*), p(\mathbf{y}))$$
$$- \Phi(\widehat{\mathrm{TP}}(\mathbf{s}',\mathbf{y}), v(\mathbf{s}'), p(\mathbf{y}))\big]$$

Let $\tilde{\mathbf{s}} = \{s_i^* \ \forall i \in [n]\setminus\{j,k\}\}$ and define $\#TP(\mathbf{z}) := \sum_i \tilde{s}_i z_i$ and $\#p(\mathbf{z}) = z_i$ (where the $\#$ prefix indicates counts rather than normalized values), and note that $v(\mathbf{s}^*) = v(\mathbf{s}')$. With these substitutions, $\mathcal{U}^L(\mathbf{s}^*;\mathbb{P}) - \mathcal{U}^L(\mathbf{s}';\mathbb{P})$ is given by:

$$\sum_{\mathbf{z}\in\{0,1\}^{n-2}} \mathbb{P}(\mathbf{z}, y_j=1, y_k=0|\mathbf{x})$$
$$\left[\Phi\left(\frac{1}{n}(\#TP(\mathbf{z})+1), v(\mathbf{s}'), \frac{1}{n}(\#p(\mathbf{z})+1)\right)\right.$$
$$\left. - \Phi\left(\frac{1}{n}\#TP(\mathbf{z}), v(\mathbf{s}'), \frac{1}{n}(\#p(\mathbf{z})+1)\right)\right]$$
$$+ \mathbb{P}(\mathbf{z}, y_j=0, y_k=1|\mathbf{x})\left[\Phi\left(\frac{1}{n}\#TP(\mathbf{z}), v(\mathbf{s}'), \frac{1}{n}(\#p(\mathbf{z})+1)\right)\right.$$
$$\left.-\Phi(\frac{1}{n}(\#TP(\mathbf{z})+1), v(\mathbf{s}'), \frac{1}{n}(\#p(\mathbf{z})+1)\right)\Bigg]$$

By conditional independence, we have that $P(\mathbf{z},y_j,y_k|\mathbf{x}) = P(\mathbf{z}|\mathbf{x})P(y_j|\mathbf{x})P(y_k|\mathbf{x})$, so that the equation further simplifies to:

$$\sum_{\mathbf{z}\in\{0,1\}^{n-2}} \mathbb{P}(\mathbf{z}|\mathbf{x})\left[\Phi\left(\frac{1}{n}(\#TP(\mathbf{z})+1), v(\mathbf{s}'),\right.\right.$$
$$\left.\left.\frac{1}{n}(\#p(\mathbf{z})+1)\right) - \Phi\left(\frac{1}{n}\#TP(\mathbf{z}), v(\mathbf{s}'), \frac{1}{n}(\#p(\mathbf{z})+1)\right)\right]$$
$$\left[\eta_j(1-\eta_k) - \eta_k(1-\eta_j)\right] =$$
$$(\eta_j-\eta_k)\sum_{\mathbf{z}\in\{0,1\}^{n-2}} \mathbb{P}(\mathbf{z}|\mathbf{x})\left[\Phi\left(\frac{1}{n}(\#TP(\mathbf{z})+1), v(\mathbf{s}'),\right.\right.$$
$$\left.\left.\frac{1}{n}(\#p(\mathbf{z})+1)\right) - \Phi\left(\frac{1}{n}\#TP(\mathbf{z}), v(\mathbf{s}'), \frac{1}{n}(\#p(\mathbf{z})+1)\right)\right]$$

Note that for each $\mathbf{z} \in \{0,1\}^{n-2}$:

- $\Phi\left(\frac{1}{n}(\#TP(\mathbf{z})+1), v(\mathbf{s}'), \frac{1}{n}(\#p(\mathbf{z})+1)\right)$ can be interpreted as L computed on the vectors $\mathbf{y} \in \mathbb{R}^n$ defined as $\{y_i = z_i \ \forall \ i \in [n]\setminus\{j,k\}\} \cup \{y_j = 1\} \cup \{y_k = 0\}$, and $\mathbf{s}^* \in \mathbb{R}^n$ (which is the assumed optimal).
- $\Phi\left(\frac{1}{n}\#TP(\mathbf{z}), v(\mathbf{s}'), \frac{1}{n}(\#p(\mathbf{z})+1)\right)$ can be interpreted as L computed on the vectors $\mathbf{y} \in \mathbb{R}^n$ defined as above and $\mathbf{s}' \in \mathbb{R}^n$.

By TP monotonicity of L, for each $\mathbf{z}$, the difference term

$$\Phi\left(\frac{1}{n}(\#TP(\mathbf{z})+1), v(\mathbf{s}'), \frac{1}{n}(\#p(\mathbf{z})+1)\right)$$
$$-\Phi\left(\frac{1}{n}\#TP(\mathbf{z}), v(\mathbf{s}'), \frac{1}{n}(\#p(\mathbf{z})+1)\right) < 0.$$

This combined with (6) implies that $\eta_j - \eta_k \geq 0$ which is a contradiction.

A.2. Proof of Theorem 2

Fix a multinomial distribution $\mathbb{P}$, and instance $\mathbf{x}$. Let the classes $C_1, C_2, \ldots, C_n$ be indexed according to the descending order of $\eta_i := \mathbb{P}(Y = C_i|\mathbf{x})$. First, observe that it suffices to show that for any fixed $0 \leq k \leq n$, the optimal solution denoted by $\mathbf{s}^*(k)$ that minimizes the expected loss restricted to subset of vectors $\mathcal{S}_k = \{\mathbf{s} \in \{0,1\}^n \mid \sum_{i=1}^n s_i = k\}$ satisfies $s_1^*(k) = s_2^*(k) = \cdots = s_k^*(k) = 1$, and $s_{k+1}^*(k) = \cdots = s_n^*(k) = 0$. Define $[[P]] = 1$ if the predicate P is true or 0 otherwise. Now, for any $\mathbf{s} \in \mathcal{S}_k$, we have,

$$\begin{aligned}
&\mathbf{E}_{Y\sim\mathbb{P}(.|\mathbf{x})}[L(\mathbf{s}, Y)] = \sum_{i\in[n]} \Phi\left(\frac{1}{n}[[s_i = 1]], \frac{k}{n}, \frac{1}{n}\right)\eta_i \\
&= \sum_{i:s_i=1} \Phi\left(\frac{1}{n}, \frac{k}{n}, \frac{1}{n}\right)\eta_i + \sum_{i:s_i=0} \Phi\left(0, \frac{k}{n}, \frac{1}{n}\right)\eta_i \\
&= \Phi\left(\frac{1}{n}, \frac{k}{n}, \frac{1}{n}\right) \sum_{i:s_i=1} \eta_i + \Phi\left(0, \frac{k}{n}, \frac{1}{n}\right)\left(1 - \sum_{i:s_i=1} \eta_i\right) \\
&= \left(\Phi\left(\frac{1}{n}, \frac{k}{n}, \frac{1}{n}\right) - \Phi\left(0, \frac{k}{n}, \frac{1}{n}\right)\right) \sum_{i:s_i=1} \eta_i \\
&\quad + \Phi\left(0, \frac{k}{n}, \frac{1}{n}\right)
\end{aligned}$$

By TP monotonicity of L, we have,

$$\Phi\left(\frac{1}{n}, \frac{k}{n}, \frac{1}{n}\right) < \Phi\left(0, \frac{k}{n}, \frac{1}{n}\right).$$

So, to minimize the RHS of the above set of equations, we need to maximize $\sum_{i:s_i=1} \eta_i$. Restricting to $\mathcal{S}_k$, the sum is maximized when we choose classes with k largest η_i values. We conclude that $\mathbf{s}^*(k)$ is the minimizer. This completes the proof.

A.3. Sufficiency of TP Monotonicity

TP monotonicity of L is sufficient but not necessary for the optimality characterization we show in the paper. For instance, consider the subclass of losses where $\Phi(\cdot, v, p)$ is independent of the first argument i.e. independent of $\widehat{\text{TP}}$. SEC is an example of a loss in this family with $\Phi_{\text{SEC}}(\widehat{\text{TP}}, v, p) = 2 - v - p$. But then, it is straight-forward to characterize optimal solution for such losses:

Proposition 5. *Let $L = \Phi(\widehat{TP}, v, p)$ be a loss independent of $\widehat{TP}$, then the optimal* (1) *under L satisfies the ordering of marginal probabilities as in Theorem 1.*

Proof. Suppose $\Phi(\cdot, v, p)$ is independent of its first argument. Let $\mathbf{s}^*$ be an optimal classifier, with $v^* = v(\mathbf{s}^*)$. If $\mathbf{s}^*$ does not already satisfy the property, then simply sort $\mathbf{s}^*$ with respect to $\mathbb{P}(Y_i|\mathbf{x})$ to obtain a new classifier $\tilde{\mathbf{s}}$. Clearly, $v(\tilde{\mathbf{s}}) = v^*$, and $\Phi(\cdot, v(\tilde{\mathbf{s}}^*), p) = \Phi(\cdot, v^*, p)$. □

A.4. Proof of Proposition 3

Suppose L satisfies TPR/TNR monotonicity. Let $u_1 = \text{TP}(\mathbf{s}_1, \mathbf{y}_1)$ and $u_2 = \text{TP}(\mathbf{s}_2, \mathbf{y}_2)$, $v = v(\mathbf{s}_1) = v(\mathbf{s}_2)$ and $p = p(\mathbf{y}_1) = p(\mathbf{y}_2)$. Note that $\Phi(u_1, v, p) = \Gamma(\frac{u_1}{p}, \frac{1-v-p+u_1}{1-p}, p)$ (and similarly equality holds for $\Phi(u_2, v, p)$). Now, whenever $u_1 = \widehat{\text{TP}}(\mathbf{s}_1, \mathbf{y}_1) > \widehat{\text{TP}}(\mathbf{s}_2, \mathbf{y}_2) = u_2$, $v(\mathbf{s}_1) = v(\mathbf{s}_2) = v$, and $p(\mathbf{y}_1) = p(\mathbf{y}_2) = p$, we have $\widehat{\text{TPR}}(\mathbf{s}_1, \mathbf{y}_1) > \widehat{\text{TPR}}(\mathbf{s}_2, \mathbf{y}_2), \widehat{\text{TNR}}(\mathbf{s}_1, \mathbf{y}_1) > \widehat{\text{TNR}}(\mathbf{s}_2, \mathbf{y}_2)$, and

$$\begin{aligned}
\Phi(u_1, v, p) &= \Gamma(\frac{u_1}{p}, \frac{1-v-p+u_1}{1-p}, p) \\
&= \Gamma(\widehat{\text{TPR}}(\mathbf{s}_1, \mathbf{y}_1), \widehat{\text{TNR}}(\mathbf{s}_1, \mathbf{y}_1), p) \\
&\overset{(*)}{<} \Gamma(\widehat{\text{TPR}}(\mathbf{s}_2, \mathbf{y}_2), \widehat{\text{TNR}}(\mathbf{s}_2, \mathbf{y}_2), p) \\
&= \Gamma(u_2.p, \frac{1-v-p+u_2}{1-p}, p) \\
&= \Phi(u_2, v, p)
\end{aligned}$$

where $(*)$ follows from TPR/TNR monotonicity of L. Thus L satisfies TP monotonicity.

B. Appendix B

B.1. Faster Algorithm for Fractional-Linear Losses

We focus our attention on the fractional-linear family of losses studied by Koyejo et al. (2014; 2015). A fractional-linear loss can be represented by Φ_{FL} as given in (4). As shown in Proposition 2, L_{FL} satisfies TP monotonicity when $c_1 < d_1$. When $c_3 = 0$ and the constants $\{d_0, d_1, d_2, d_3\}$ are rational in (4), we can get a quadratic-time procedure for computing $\mathbf{s}^*$ appealing to the method proposed by Ye et al. (2012). Formally, we consider the sub-family of TP monotonic fractional-linear losses:

$$\{L_{SFL} : \Phi_{FL}(u, v, p) = \frac{c_0 + c_1 u + c_2 v}{d_0 + d_1 u + d_2 v + d_3 p},\ c_1 < d_1, \text{ and } d_0, d_1, d_2, d_3 \text{ are rational}\}, \tag{7}$$

which includes the loss based on Jaccard measure and others not studied by Ye et al. (2012). Consider Step 6 of Algorithm 1 for a loss in family (7):

$$L_k \leftarrow \sum_{0\leq k_1\leq k} C[k_1](c_0 n + c_1 k_1 + c_2 k)\ .$$

$$\sum_{0\leq k_2\leq n-k} D[k_2]/(d_0 n + (d_1 + d_3)k_1 + d_2 k + d_3 k_2).$$

Define $b(k, \alpha) = \sum_{0\leq k_2\leq n-k} D_k[k_2]/(\alpha + d_3 k_2)$. Verify that $b(n, \alpha) = 1/\alpha$. From the fact that $D_{k-1}[i] = \eta_k D_k[i-1] + (1-\eta_k)D_k[i]$, it follows that:

$$b(k-1, \alpha) = \eta_k b(k, \alpha + d_3) + (1 - p_k)b(k, \alpha).$$

Now, when d_i's are rational, i.e. $d_i = q_i/r_i$, the above induction can be implemented using an array to store the values of b, for possible values of α.

Algorithm 2 Computing $\mathbf{s}^*$ for L_{SFL} in the family (7)

1: **Input:** Estimates of $\eta_i = \mathbb{P}(Y_i = 1|\mathbf{x}), i = 1, 2, \ldots, n$ sorted wrt. η_i, and $c_0, c_1, c_2, d_i = q_i/r_i, i = 0, 1, 2, 3$ corresponding to L_{SFL}
2: Init $s_i^* = 0, \forall i \in [n]$.
3: Set $j_0 \leftarrow r_1 r_2 r_3 q_0$, $j_{u,1} \leftarrow r_0 r_2 r_3 q_1$, $j_{u,2} \leftarrow r_0 r_1 r_2 q_3$, $j_v \leftarrow r_0 r_1 r_3 q_2$
4: **for** $1 \le i \le (|j_{u,1}| + |j_{u,2}| + |j_v|)n$ **do**
5: set $S[i] \leftarrow r_0 r_1 r_2 r_3/(i + j_0 n)$.
6: **end for**
7: **for** $k = n$ to 1 **do**
8: For $0 \le i \le k$, set $C_k[i]$ as the coefficient of z^i in $\Pi_{i=1}^{k}\big(\eta_i z + (1 - \eta_i)\big)$.
9: $L_{\text{SFL};k} \leftarrow \sum_{0 \le k_1 \le k} (c_0 n + c_1 k_1 + c_2 k) C_k[k_1] S[(j_{u,1} + j_{u,2})k_1 + j_v k]$.
10: **for** $i = 1$ to $(|j_{u,1}| + |j_{u,2}| + |j_v|)(k - 1)$ **do**
11: $S[i] \leftarrow (1 - \eta_k) S[i] + \eta_k S[i + j_{u,2}]$.
12: **end for**
13: **end for**
14: Set $k^* \leftarrow \arg\min_k L_{\text{SFL};k}$ and $s_i^* \leftarrow 1$ for $i \in [k^*]$.
15: return $\mathbf{s}^*$

Correctness of Algorithm 2: When $d_3 \neq 0$, at line 7 of Algorithm 2, we can verify that $S[i] = b(k, (i + j_0 n)d_3/j_{u,2})$, and therefore at line 9, $S[(j_{u,1} + j_{u,2})k_1 + j_v k] = b(k, (j_{u,1} + j_{u,2})k_1 + j_v k + j_0 n)d_3/j_{u,2}) = b(k, (d_1 + d_3)k_1 + d_2 k + d_0 n)$ as desired. When $d_3 = 0$, $b(k, \alpha) = b(k - 1, \alpha)$ for all $1 \le k \le n$. Let $q_3 = 0$ and $r_3 = 1$. Then, line 5 sets $S[i] = r_0 r_1 r_2/(i + j_0 n)$, line 11 maintains this invariant as $j_{u,2} = 0$ in this case, and therefore at line 9, $S[(j_{u,1} + j_{u,2})k_1 + j_v k] = 1/(d_1 k_1 + d_2 k + d_0 n)$ as desired.

B.2. Proof of Theorem 3

For convenience, define:

$$\mathcal{U}^L(\mathbf{s}; \mathbb{P}) := \mathbf{E}_{\mathbf{Y} \sim \mathbb{P}(.|\mathbf{x})} L(\mathbf{s}, \mathbf{Y}) \, .$$

Let $\mathcal{U}_*^L := \mathcal{U}^L(\mathbf{s}^*; \mathbb{P})$ and let $\widehat{\mathcal{U}}^L = \mathcal{U}^L(\hat{\mathbf{s}}; \mathbb{P})$. Also define the empirical distribution:

$$\hat{\mathbb{P}}(\mathbf{y}|\mathbf{x}) = \Pi_{i=1}^{n} \hat{\eta}_i^{y_i} (1 - \hat{\eta}_i)^{1 - y_i}.$$

Now consider:

$$\begin{aligned} \widehat{\mathcal{U}}^L - \mathcal{U}_*^L = \quad & \widehat{\mathcal{U}}^L + \mathcal{U}^L(\hat{\mathbf{s}}; \hat{\mathbb{P}}) - \mathcal{U}^L(\hat{\mathbf{s}}; \hat{\mathbb{P}}) - \mathcal{U}_*^L \\ \le \quad & \widehat{\mathcal{U}}^L + \mathcal{U}^L(\mathbf{s}^*; \hat{\mathbb{P}}) - \mathcal{U}^L(\hat{\mathbf{s}}; \hat{\mathbb{P}}) - \mathcal{U}_*^L \\ \le \quad & 2 \max_{\mathbf{s}} \big| \mathcal{U}^L(\mathbf{s}; \mathbb{P}) - \mathcal{U}^L(\mathbf{s}; \hat{\mathbb{P}}) \big| \end{aligned} \tag{8}$$

For any fixed $\mathbf{s} \in \{0, 1\}^n$, we have:

$$\begin{aligned} & \big| \mathcal{U}^L(\mathbf{s}; \mathbb{P}) - \mathcal{U}^L(\mathbf{s}; \hat{\mathbb{P}}) \big| = \\ & \Big| \sum_{y \in \{0,1\}^n} \hat{\mathbb{P}}(\mathbf{y}|\mathbf{x}) L(\mathbf{s}, \mathbf{y}) - \sum_{y \in \{0,1\}^n} \mathbb{P}(\mathbf{y}|\mathbf{x}) L(\mathbf{s}, \mathbf{y}) \Big| \\ & \le \sum_{y \in \{0,1\}^n} \big| \hat{\mathbb{P}}(\mathbf{y}|\mathbf{x}) - \mathbb{P}(\mathbf{y}|\mathbf{x}) \big| L(\mathbf{s}, \mathbf{y}) \end{aligned} \tag{9}$$

Let $\eta(x)$ denote the empirical estimate obtained using m training samples. Now because $\hat{\eta}(x) \xrightarrow{p} \eta(x)$, we have that for sufficiently large set of training examples, $\hat{\mathbb{P}}(\mathbf{y}|\mathbf{x}) \xrightarrow{p} \mathbb{P}(\mathbf{y}|\mathbf{x})$; i.e. for any given $\epsilon > 0$, there exists m_ϵ such that for all $m > m_\epsilon$, $|\hat{\mathbb{P}}(\mathbf{y}|\mathbf{x}) - \mathbb{P}(\mathbf{y}|\mathbf{x})| < \epsilon$, with high probability. It follows that, with high probability, (9) $\le \epsilon \sum_{y \in \{0,1\}^n} L(\mathbf{s}, \mathbf{y})$. Assuming L is bounded, we have that for any fixed s, $\big| \mathcal{U}^L(\mathbf{s}; \mathbb{P}) - \mathcal{U}^L(\mathbf{s}; \hat{\mathbb{P}}) \big| \le C\epsilon$, for some constant C that depends only on the metric L and (fixed) test set size n. The uniform convergence also follows because the max in (8) is over finitely many vectors $\mathbf{s}$. Putting together, we have that for any given $\delta, \epsilon' > 0$, there exists training sample size $m_{\epsilon',\delta}$ such that the output $\hat{\mathbf{s}}$ of our procedure satisfies, with probability at least $1 - \delta$, $\widehat{\mathcal{U}}^L - \mathcal{U}_*^L < \epsilon'$; when L is unbounded, we have that $\mathbf{s}^* = \arg\min_{\mathbf{s} \in \{0,1\}^n} L(\mathbf{s}, \cdot)$ over all unbounded $L(\mathbf{s}, \mathbf{y})$. Thus all that is required is support consistency i.e. $\{\mathbf{y}|\hat{\mathbb{P}}(\mathbf{y}|\mathbf{x}) > 0\} \xrightarrow{p} \{\mathbf{y}|\mathbb{P}(\mathbf{y}|\mathbf{x}) > 0\}$ which is a much weaker condition than distribution consistency. The proof is complete.

C. Appendix C

EUM and DTA Classification

A recent flurry of theoretical results and practical algorithms highlights a growing interest in understanding and optimizing non-decomposable metrics (Dembczynski et al., 2011; Ye et al., 2012; Koyejo et al., 2014; Narasimhan et al., 2014). Existing theoretical analysis has focused on two distinct approaches for characterizing the *population* version of the non-decomposable metrics: identified by Ye et al. (2012) as decision theoretic analysis (DTA) and empirical utility maximization (EUM). DTA population utilities measure the expected gain of a classifier on a fixed-size test set, while EUM population utilities are a function of the population confusion matrix. In other words, DTA population utilities measure the the average utility over an infinite set of test sets, each of a fixed size, while EUM population utilities evaluate the performance of a classifier over a single infinitely large test set.

It has recently been shown that for EUM based population utilities, the optimal classifier for large classes of non-decomposable binary classification metrics is just the sign of the thresholded conditional probability of the posi-

tive class with a metric-dependent threshold (Koyejo et al., 2014; Narasimhan et al., 2014). In addition, practical algorithms have been proposed for such EUM consistent classification based on direct optimization for the threshold on a held-out validation set. In stark contrast to this burgeoning understanding of EUM optimal classification, we are aware of only two metrics for which DTA consistent classifiers have been derived and shown to exhibit a simple form; namely, the F_β metric (Lewis, 1995; Dembczynski et al., 2011; Ye et al., 2012) and squared error in counting (SEC) studied by Lewis (1995).

While the optimal classifiers of both EUM and DTA population utilities associated with the performance metrics we study comprise signed thresholding of the conditional probability of the positive class, the evaluation and optimization for EUM and DTA utilities require quite different techniques. Given a classifier and a distribution, evaluating a population DTA utility can involve exponential-time computation, even leaving aside maximizing the utility on a fixed test set. As we show, in light of the probability ranking principle, and with careful implementation, this can actually be reduced to cubic complexity. These computations can be further reduced to quadratic complexity in a few special cases (Ye et al., 2012). To this end, we propose two algorithms for optimal DTA classification. The first algorithm runs in $O(n^3)$ time for a general metric, where n is the size of the test set and the second algorithm runs in time $O(n^2)$ for special cases such as F_β and Jaccard. We show that our overall procedure for decision-theoretic classification is consistent. More recently, Parambath et al. (2014) gave a theoretical analysis of the binary and multi-label F_β measure in the EUM setting. Dembczynski et al. (2011) analyzed the F_β measure in the DTA setting including the case where the data is non i.i.d., and also proposed efficient algorithms for optimal classification.