# Consistent Multiclass Algorithms for Complex Performance Measures

# Supplementary Material

**Notations.** Let $\lambda$ be the base measure over $\Delta_n$ given by the uniform random variable (say $U$) over $\Delta_n$. Hence, for all measurable $\mathcal{A} \subseteq \Delta_n$, $\lambda(\mathcal{A}) = \mathbf{P}(U \in \mathcal{A})$. Also, $\boldsymbol{\eta} : \mathcal{X} \to \Delta_n$ is the mapping that gives the conditional probability vector $\boldsymbol{\eta}(x) = [\mathbf{P}(Y = 1 \,|\, X = x), \ldots, \mathbf{P}(Y = n \,|\, X = x)]^\top \in \Delta_n$ for a given instance $x \in \mathcal{X}$. Let $\nu$ be the probability measure over the simplex induced by the random variable $\boldsymbol{\eta}(X)$; in particular, for all measurable $\mathcal{A} \subseteq \Delta_n$, $\nu(\mathcal{A}) = \mathbf{P}_{X \sim \mu}(\boldsymbol{\eta}(X) \in \mathcal{A})$. For a matrix $\mathbf{L} \in [0,1]^{n \times n}$ we let $\boldsymbol{\ell}_1, \boldsymbol{\ell}_2, \ldots, \boldsymbol{\ell}_n$ be the columns of $\mathbf{L}$. For any set $\mathcal{A} \subseteq \mathbb{R}^d$, the set $\overline{\mathcal{A}}$ denotes its closure. For any vector $\mathbf{v} \in \mathbb{R}^n$, we let $(\mathbf{v})_{(i)}$ denote the $i^{th}$ element when the components of $\mathbf{v}$ are sorted in ascending order. For any $y \in [n]$, we shall denote $\text{rand}(y) = [\mathbf{1}(y = 1), \ldots, \mathbf{1}(y = n)]^\top$.

## A. Supplementary Material For Section 2 (Complex Performance Measures)

### A.1. Details of Micro $F_1$-measure

We consider the form of the micro $F_1$ used in the BioNLP challenge (Kim et al., 2013), which treats class 1 as a 'default' class (in information extraction, this class pertains to examples for which no information is required to be extracted). One can then define the micro precision of a classifier $h : \mathcal{X} \to [n]$ with confusion matrix $\mathbf{C} = \mathbf{C}^D[h]$ as the probability of an instance being correctly labelled, given that it was assigned by $h$ a class other than 1:

$$\text{microPrec}(\mathbf{C}) \;=\; \mathbf{P}(h(X) = Y \,|\, h(X) \neq 1) \;=\; \frac{\sum_{i=2}^n C_{ii}}{\sum_{i=2}^n \sum_{j=1}^n C_{ji}} \;=\; \frac{\sum_{i=2}^n C_{ii}}{1 - \sum_{i=1}^n C_{i1}}$$

Similarly, the micro recall of $h$ can be defined as the probability of an instance being correctly labelled, given that its true class was not 1:

$$\text{microRec}(\mathbf{C}) \;=\; \mathbf{P}(h(X) = Y \,|\, Y \neq 1) \;=\; \frac{\sum_{i=2}^n C_{ii}}{\sum_{i=2}^n \sum_{j=1}^n C_{ij}} \;=\; \frac{\sum_{i=2}^n C_{ii}}{1 - \sum_{i=1}^n C_{1i}}.$$

The micro $F_1$ that we analyze is the harmonic mean of the micro precision and micro recall given above:

$$\psi^{\text{micro}F_1}(\mathbf{C}) \;=\; \frac{2 \times \text{microPrec}(\mathbf{C}) \times \text{microRec}(\mathbf{C})}{\text{microPrec}(\mathbf{C}) \;+\; \text{microRec}(\mathbf{C})} \;=\; \frac{2 \sum_{i=2}^n C_{ii}}{2 - \sum_{i=1}^n C_{1i} - \sum_{i=1}^n C_{i1}}\,.$$

Note that the above performance measure can be written as a ratio-of-linear function: $\psi^{\text{micro}F_1}(\mathbf{C}) = \frac{\langle \mathbf{A}, \mathbf{C} \rangle}{\langle \mathbf{B}, \mathbf{C} \rangle}$, where $A_{11} = 0, A_{ii} = 2, \forall i \neq 1, A_{ij} = 0, \forall i \neq j$, and $B_{11} = 0, B_{1i} = B_{i1} = 1, \forall i \neq 1, B_{ij} = 2, \forall i \neq j$. Also, note that this performance measure satisfies the condition in Theorem 17 with $\sup_{\mathbf{C} \in \mathcal{C}_D} \psi^{\text{micro}F_1}(\mathbf{C}) \leq 1$, and $\min_{\mathbf{C} \in \mathcal{C}_D} \langle \mathbf{B}, \mathbf{C} \rangle \geq 1 - \pi_1 > 0$, and hence Algorithm 2 is consistent for this performance measure.

Recently, Parambath et al. (2014) also considered a form of micro $F_1$ similar to that used in the BioNLP challenge. The expression they use is slightly simpler than ours and differs slightly from the BioNLP performance measure:

$$\psi^{\text{micro}F_1}(\mathbf{C}) = \frac{2 \sum_{i=2}^n C_{ii}}{1 + \sum_{i=2}^n C_{ii} - C_{11}}\,.$$

Another popular variant of the micro $F_1$ involves averaging the entries of the 'one-versus-all' binary confusion matrices for all classes, and computing the $F_1$ for the averaged matrix; as pointed out by Manning et al. (2008), this form of micro $F_1$ effectively reduces to the 0-1 classification accuracy.

## B. Supplementary Material for Section 3 (Bayes Optimal Classifiers)

### B.1. Example Distribution Where the Optimal Classifier Needs to be Randomized

We present an example distribution where the optimal performance for the G-mean measure can be achieved only by a randomized classifier.

**Example 5** (Distribution where the optimal classifier needs to be randomized)**.** *Let $D$ be a distribution over $\{x\} \times \{1, 2\}$ with $\eta_1(x) = \eta_2(x) = \frac{1}{2}$ and suppose we are interested in finding the optimal classifier for the G-mean performance measure (see Example 3) under $D$. The two deterministic classifiers for this setting, namely, one which predicts 1 on $x$ and the other that predicts 2 on $x$, yield a G-mean of 0. However, the randomized classifier $h^*(x) = \left[\frac{1}{2}, \frac{1}{2}\right]^\top$ has a G-mean value of $\frac{1}{4} > 0$ and can be verified to be the unique optimal classifier for G-mean under $D$.*

We next present the proofs for the theorems/lemmas/propositions in Section 3.

### B.2. Proof of Theorem 11

**Theorem** (Form of Bayes optimal classifier for ratio-of-linear $\psi$)**.** *Let $\psi : [0, 1]^{n\times n} \to \mathbb{R}_+$ be a ratio-of-linear performance measure of the form $\psi(\mathbf{C}) = \frac{\langle \mathbf{A}, \mathbf{C}\rangle}{\langle \mathbf{B}, \mathbf{C}\rangle}$ for some $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times n}$ with $\langle \mathbf{B}, \mathbf{C}\rangle > 0 \ \forall \mathbf{C} \in \mathcal{C}_D$. Let $t^*_D = \mathcal{P}^{\psi,*}_D$. Let $\widetilde{\mathbf{L}}^* = -(\mathbf{A} - t^*_D\mathbf{B})$, and let $\mathbf{L}^* \in [0, 1]^{n\times n}$ be obtained by scaling and shifting $\widetilde{\mathbf{L}}^*$ so its entries lie in $[0, 1]$. Then any classifier that is $\psi^{\mathbf{L}^*}$-optimal is also $\psi$-optimal.*

In the following, we omit the subscript on $t^*_D$ for easy of presentation. We first state the following lemma using which we prove the above theorem.

**Lemma 18.** *Let $\psi : [0, 1]^{n\times n} \to \mathbb{R}_+$ be such that $\psi(\mathbf{C}) = \frac{\langle \mathbf{A}, \mathbf{C}\rangle}{\langle \mathbf{B}, \mathbf{C}\rangle}$, for some matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times n}$ with $\langle \mathbf{B}, \mathbf{C}\rangle > 0$ for all $\mathbf{C} \in \mathcal{C}_D$. Let $t^* = \sup_{\mathbf{C}\in\mathcal{C}_D} \psi(\mathbf{C})$. Then $\sup_{\mathbf{C}\in\mathcal{C}_D} \langle \mathbf{A} - t^*\mathbf{B}, \mathbf{C}\rangle = 0$.*

*Proof.* Define $\phi : \mathbb{R} \to \mathbb{R}$ as $\phi(t) = \sup_{\mathbf{C}\in\mathcal{C}_D} \langle \mathbf{A} - t\mathbf{B}, \mathbf{C}\rangle$. It is easy to see that $\phi$ (being a point-wise supremum of linear functions) is convex, and hence a continuous function over $\mathbb{R}$. By definition of $t^*$, we have for all $\mathbf{C} \in \mathcal{C}_D$,

$$\frac{\langle \mathbf{A}, \mathbf{C}\rangle}{\langle \mathbf{B}, \mathbf{C}\rangle} \leq t^* \quad \text{or equivalently} \quad \phi(t^*) = \langle \mathbf{A} - t^*\mathbf{B}, \mathbf{C}\rangle \leq 0.$$

Thus

$$\phi(t^*) = \sup_{\mathbf{C}\in\mathcal{C}_D} \langle \mathbf{A} - t^*\mathbf{B}, \mathbf{C}\rangle \leq 0 \,. \tag{2}$$

Also, for any $t < t^*$, there exists $\mathbf{C} \in \mathcal{C}_D$ such that

$$\frac{\langle \mathbf{A}, \mathbf{C}\rangle}{\langle \mathbf{B}, \mathbf{C}\rangle} > t \quad \text{or equivalently} \quad \phi(t) = \langle \mathbf{A} - t\mathbf{B}, \mathbf{C}\rangle > 0.$$

Thus for all $t < t^*$,

$$\phi(t) = \sup_{\mathbf{C}\in\mathcal{C}_D} \langle \mathbf{A} - t\mathbf{B}, \mathbf{C}\rangle > 0 \,.$$

Next, by continuity of $\phi$, for any monotonically increasing sequence of real numbers $\{t_i\}_{i=1}^\infty$ converging to $t^*$, we have that $\phi(t_i)$ converges to $\phi(t^*)$; since for each $t_i$ in this sequence $\phi(t_i) > 0$, at the $t^*$, we have that $\phi(t^*) \geq 0$. Along with Eq. (2), this gives us

$$\sup_{\mathbf{C}\in\mathcal{C}_D} \langle \mathbf{A} - t^*\mathbf{B}, \mathbf{C}\rangle = \phi(t^*) = 0.$$

□

We next give the proof for Theorem 11

*Proof of Theorem 11.* Let $h^* : \mathcal{X} \to \Delta_n$ be a $\psi^{\mathbf{L}^*}$-optimal classifier. We shall show that $h^*$ is also $\psi$-optimal, which will also imply existence of the $\psi$-optimal classifier. Then we have

$$1 - \langle \mathbf{L}^*, \mathbf{C}^D[h^*]\rangle = \sup_{h:\mathcal{X}\to\Delta_n} 1 - \langle \mathbf{L}^*, \mathbf{C}^D[h]\rangle = \sup_{\mathbf{C}\in\mathcal{C}_D} 1 - \langle \mathbf{L}^*, \mathbf{C}\rangle.$$

Since $\mathbf{L}^*$ is a scaled and translated version of $\widetilde{\mathbf{L}}^* = \mathbf{A} - t^*\mathbf{B}$ (where $t^* = \mathcal{P}^{\psi,*}$), we further have

$$\langle \mathbf{A} - t^*\mathbf{B}, \mathbf{C}^D[h^*]\rangle = \sup_{\mathbf{C}\in\mathcal{C}_D} \langle \mathbf{A} - t^*\mathbf{B}, \mathbf{C}\rangle.$$

Now, from Lemma 18 we know that $\langle \mathbf{A} - t^*\mathbf{B}, \mathbf{C}^D[h^*]\rangle = 0$. Hence,

$$\frac{\langle \mathbf{A}, \mathbf{C}^D[h^*]\rangle}{\langle \mathbf{B}, \mathbf{C}^D[h^*]\rangle} = t^*,$$

or equivalently,

$$\psi(\mathbf{C}^D[h^*]) = \sup_{\mathbf{C}\in\mathcal{C}_D} \psi(\mathbf{C}).$$

Thus $h^*$ is also $\psi$-optimal, which completes the proof. □

### B.3. Proof of Proposition 10

**Proposition.** *$\mathcal{C}_D$ is a convex set.*

*Proof.* Let $\mathbf{C}_1, \mathbf{C}_2 \in \mathcal{C}_D$. Let $\lambda \in [0, 1]$. We will show that $\lambda\mathbf{C}_1 + (1-\lambda)\mathbf{C}_2 \in \mathcal{C}_D$.

By definition of $\mathcal{C}_D$, there exists randomized classifiers $h_1, h_2 : \mathcal{X}\to\Delta_n$ such that $\mathbf{C}_1 = \mathbf{C}^D[h_1]$ and $\mathbf{C}_2 = \mathbf{C}^D[h_2]$.

Consider the randomized classifier $h^\lambda : \mathcal{X}\to\Delta_n$ defined as

$$h^\lambda(x) = \lambda h_1(x) + (1-\lambda)h_2(x)\ .$$

It can be seen that

$$\mathbf{C}^D[h^\lambda] = \lambda\mathbf{C}_1 + (1-\lambda)\mathbf{C}_2\ .$$

□

### B.4. Supporting Technical Lemmas For Lemma 12 and Theorem13

In this subsection we give some supporting technical lemmas which will be useful in the proofs for Lemma 12 and Theorem 13.

**Lemma 19 (Confusion matrix as an integration).** *Let $\mathbf{f} : \Delta_n\to\Delta_n$. Then*

$$\mathbf{C}^D[\mathbf{f}\circ\boldsymbol{\eta}] = \int_{\mathbf{p}\in\Delta_n} \mathbf{p}(\mathbf{f}(\mathbf{p}))^\top d\nu(\mathbf{p})\ .$$

*Proof.*

$$\begin{aligned} C^D_{i,j}[\mathbf{f}\circ\boldsymbol{\eta}] &= \mathbf{E}_{(X,Y)\sim D}[f_j(\boldsymbol{\eta}(X))\cdot\mathbf{1}(Y=i)] \\ &= \mathbf{E}_{\mathbf{p}\sim\nu}\mathbf{E}_{(X,Y)\sim D}\left[f_j(\mathbf{p})\cdot\mathbf{1}(Y=i)\middle|\boldsymbol{\eta}(X)=\mathbf{p}\right] \\ &= \mathbf{E}_{\mathbf{p}\sim\nu}\left[p_i\cdot f_j(\mathbf{p})\right]. \end{aligned}$$

□

**Proposition 20 (Sufficiency of conditional probability).** *Let $D$ be a distribution over $\mathcal{X}\times\mathcal{Y}$. For any randomized classifier $h : \mathcal{X}\to\Delta_n$ there exists another randomized classifier $h' : \mathcal{X}\to\Delta_n$ such that $\mathbf{C}^D[h] = \mathbf{C}^D[h']$ and $h'$ is such that $h' = \mathbf{f}\circ\boldsymbol{\eta}$, for some $\mathbf{f} : \Delta_n\to\Delta_n$.*

*Proof.* Let $h : \mathcal{X}\to\Delta_n$. Define $\mathbf{f} : \Delta_n\to\Delta_n$ as follows,

$$\mathbf{f}(\mathbf{p}) = \mathbf{E}_{X\sim\mu}[h(X)|\boldsymbol{\eta}(X)=\mathbf{p}]\ .$$

We then have for any $i, j \in [n]$ that,

$$\begin{aligned} C^D_{i,j}[h] &= \mathbf{E}_{(X,Y)\sim D}[h_j(X)\cdot\mathbf{1}(Y=i)] \\ &= \mathbf{E}_{\mathbf{p}\sim\nu}\mathbf{E}_{(X,Y)\sim D}[h_j(X)\cdot\mathbf{1}(Y=i)|\boldsymbol{\eta}(X)=\mathbf{p}] \end{aligned}$$

$$\begin{aligned} &= \mathbf{E}_{\mathbf{p}\sim\nu}\left[\mathbf{E}_{(X,Y)\sim D}[h_j(X)|\boldsymbol{\eta}(X)=\mathbf{p}]\cdot\mathbf{E}_{(X,Y)\sim D}[\mathbf{1}(Y=i)|\boldsymbol{\eta}(X)=\mathbf{p}]\right] \\ &= \mathbf{E}_{\mathbf{p}\sim\nu}\left[f_j(\mathbf{p})\cdot p_i\right] \\ &= C^D_{i,j}[\mathbf{f}\circ\boldsymbol{\eta}] \end{aligned}$$

where the third equality follows because, given $\boldsymbol{\eta}(X)$, the random variables $X$ and $Y$ are independent, and the last inequality follows from Lemma 19. □

**Lemma 21 (Continuity of the $\mathbf{C}^D$ mapping).** *Let $D$ be a distribution over $\mathcal{X}\times\mathcal{Y}$. Let $\mathbf{f}_1,\mathbf{f}_2:\Delta_n\to\Delta_n$. Then*

$$\left|\left|\mathbf{C}^D[\mathbf{f}_1\circ\boldsymbol{\eta}]-\mathbf{C}^D[\mathbf{f}_2\circ\boldsymbol{\eta}]\right|\right|_1 \leq \int_{\mathbf{p}\in\Delta_n}||\mathbf{f}_1(\mathbf{p})-\mathbf{f}_2(\mathbf{p})||_1 d\nu(\mathbf{p})\ .$$

*Proof.* Let $\mathbf{f}_1,\mathbf{f}_2:\Delta_n\to\Delta_n$

$$\begin{aligned} \mathbf{C}^D[\mathbf{f}_1\circ\boldsymbol{\eta}]-\mathbf{C}^D[\mathbf{f}_2\circ\boldsymbol{\eta}] &= \int_{\mathbf{p}\in\Delta_n}\mathbf{p}(\mathbf{f}_1(\mathbf{p})-\mathbf{f}_2(\mathbf{p}))^\top d\nu(\mathbf{p}) \\ \left|\left|\mathbf{C}^D[\mathbf{f}_1\circ\boldsymbol{\eta}]-\mathbf{C}^D[\mathbf{f}_2\circ\boldsymbol{\eta}]\right|\right|_1 &\leq \int_{\mathbf{p}\in\Delta_n}||\mathbf{p}(\mathbf{f}_1(\mathbf{p})-\mathbf{f}_2(\mathbf{p}))^\top||_1 d\nu(\mathbf{p}) \\ &= \int_{\mathbf{p}\in\Delta_n}||\mathbf{p}||_1||\mathbf{f}_1(\mathbf{p})-\mathbf{f}_2(\mathbf{p})||_1 d\nu(\mathbf{p}) \\ &= \int_{\mathbf{p}\in\Delta_n}||\mathbf{f}_1(\mathbf{p})-\mathbf{f}_2(\mathbf{p})||_1 d\nu(\mathbf{p}). \end{aligned}$$

□

**Lemma 22 (Volume of a inverse linear map of an interval).** *Let $d>0$ be any integer. Let $\mathcal{V}\subseteq\mathbb{R}^d$ be compact and convex. Let $f:\mathbb{R}^d\to\mathbb{R}$ be an affine function such that it is non-constant over $\mathcal{V}$. Let $V$ be a vector valued random variable taking values uniformly over $\mathcal{V}$. Then, there exists a constant $\alpha>0$ such that for all $c\in\mathbb{R}$ and $\epsilon\in\mathbb{R}_+$ we have*

$$\mathbf{P}(f(V)\in[c,c+\epsilon])\leq\alpha\epsilon\ .$$

*Proof.* Let us assume for now that affine hull of $\mathcal{V}$ is the entire space $\mathbb{R}^d$.

For any integer $i$ and set $\mathcal{A}$, let $\text{vol}_i(\mathcal{A})$ denote the $i$-th dimensional volume of the set $\mathcal{A}$. Note that $\text{vol}_i(\mathcal{A})$ is undefined if the affine-hull dimension of $\mathcal{A}$ is greater than $i$ and is equal to zero if the affine-hull dimension of $\mathcal{A}$ is lesser than $i$.

For any $r>0$ and any integer $i>0$ let $B_i(r)\subseteq\mathbb{R}^i$ denote the set $B_i(r)=\{\mathbf{x}\in\mathbb{R}^i:||\mathbf{x}||_2\leq r\}$. Also let $R$ be the smallest value such that $\mathcal{V}\subseteq B_d(R)$.

Let the affine function $f$ be such that for all $\mathbf{x}\in\mathbb{R}^d$, the value $f(\mathbf{x})=\mathbf{g}^\top\mathbf{x}+u$. By the assumption of non-constancy of $f$ on $\mathcal{V}$ we have that $\mathbf{g}\neq 0$.

We now have that

$$\begin{aligned} \mathbf{P}(f(V)\in[c,c+\epsilon]) &= \frac{\text{vol}_d\left(\{\mathbf{v}\in\mathcal{V}:c-u\leq\mathbf{g}^\top\mathbf{v}\leq c-u+\epsilon\}\right)}{\text{vol}_d(\mathcal{V})} \\ &\leq \frac{\text{vol}_d\left(\{\mathbf{v}\in B_d(R):c-u\leq\mathbf{g}^\top\mathbf{v}\leq c-u+\epsilon\}\right)}{\text{vol}_d(\mathcal{V})} \\ &\leq \epsilon\cdot\frac{\text{vol}_{d-1}\left(B_{d-1}(R)\right)}{\text{vol}_d(\mathcal{V})||\mathbf{g}||_2}\ . \end{aligned}$$

The last inequality follows from the observation that $d$-volume of a strip of a $d$ dimensional sphere of radius $r$ is at most the $d-1$ volume of a $d-1$ dimensional sphere of radius $r$ times the width of the strip, and the width of the strip under consideration here is simply $\frac{\epsilon}{||\mathbf{g}||_2}$.

Finally, if the affine hull of $\mathcal{V}$ is not the entire space $\mathbb{R}^d$, one can simply consider the affine-hull of $\mathcal{V}$ to be the entire (lesser dimensional) space and all the above arguments hold with some affine transformations and a smaller $d$. □

**Lemma 23 (Fraction of instances with the best and second best prediction being similar in performance is small).** *Let $\mathbf{L} \in [0,1]^{n\times n}$ be such that no two columns are identical. Let the distribution $D$ over $\mathcal{X} \times \mathcal{Y}$ be such that the measure over conditional probabilities $\nu$, is absolutely continuous w.r.t. the base measure $\lambda$. Let $c \geq 0$. Let $\mathcal{A}_c \subseteq \Delta_n$ be the set*

$$\mathcal{A}_c = \{\mathbf{p} \in \Delta_n : (\mathbf{p}^\top \mathbf{L})_{(2)} - (\mathbf{p}^\top \mathbf{L})_{(1)} \leq c\}$$

*Let $r : \mathbb{R}_+ \to \mathbb{R}_+$ be the function defined as*

$$r(c) = \nu(\mathcal{A}_c) \ .$$

*Then*

*(a) $r$ is a monotonically increasing function.*

*(b) There exists a $C > 0$ such that $r$ is a continuous function over $[0, C]$.*

*(c) $r(0) = 0$.*

*Proof. Part (a):*

The fact that $r$ is a monotonically increasing function is immediately obvious from the observation that $\mathcal{A}_a \subseteq \mathcal{A}_b$ for any $a < b$.

*Part (b):*

Let

$$C = \frac{1}{2} \min\{d \in \mathbb{R} : \boldsymbol{\ell}_y - \boldsymbol{\ell}_{y'} = d\mathbf{e} \text{ for some } y, y' \in [n], y \neq y'\} \ ,$$

where $\mathbf{e}$ is the all ones vector. If there exists no $y, y'$ such that $\boldsymbol{\ell}_y - \boldsymbol{\ell}_{y'}$ is a scalar multiple of $\mathbf{e}$, then we simply set $C = \infty$. Note that by our assumption on unequal columns on $\mathbf{L}$, we always have $C > 0$.

For any $c > 0$ and $y, y' \in [n]$ with $y \neq y'$, define the set $\mathcal{A}_c^{y,y'}$ as

$$\mathcal{A}_c^{y,y'} = \{\mathbf{p} \in \Delta_n : \mathbf{p}^\top \boldsymbol{\ell}_y - \mathbf{p}^\top \boldsymbol{\ell}_{y'} \leq c\} \ .$$

For any $c, \epsilon > 0$, it can be clearly seen that

$$\begin{aligned}
\nu(\mathcal{A}_{c+\epsilon}) - \nu(\mathcal{A}_c) &= \nu(\mathcal{A}_{c+\epsilon} \setminus \mathcal{A}_c) \ , \\
\mathcal{A}_{c+\epsilon} \setminus \mathcal{A}_c &\subseteq \bigcup_{y,y' \in [n], y \neq y'} \left(\mathcal{A}_{c+\epsilon}^{y,y'} \setminus \mathcal{A}_c^{y,y'}\right) \ , \\
\nu(\mathcal{A}_{c+\epsilon} \setminus \mathcal{A}_c) &\leq \sum_{y,y' \in [n], y \neq y'} \nu\left(\mathcal{A}_{c+\epsilon}^{y,y'} \setminus \mathcal{A}_c^{y,y'}\right) \ .
\end{aligned}$$

Hence, our proof for continuity of $r$ would be complete, if we show that $\nu\left(\mathcal{A}_{c+\epsilon}^{y,y'} \setminus \mathcal{A}_c^{y,y'}\right)$ goes to zero as $\epsilon$ goes to zero for all $y \neq y'$ and $c \in [0, C]$.

Let $c \in [0, C]$ and $y, y' \in [n]$ with $y \neq y'$

$$\mathcal{A}_{c+\epsilon}^{y,y'} \setminus \mathcal{A}_c^{y,y'} = \{\mathbf{p} \in \Delta_n : c < \mathbf{p}^\top (\boldsymbol{\ell}_y - \boldsymbol{\ell}_{y'}) \leq c + \epsilon\} \ .$$

If $\boldsymbol{\ell}_y - \boldsymbol{\ell}_{y'} = d\mathbf{e}$ for some $d$, we have that $\mathbf{p}^\top (\boldsymbol{\ell}_y - \boldsymbol{\ell}_{y'}) = d$ and $d > C$ by definition of $C$. Hence for small enough $\epsilon$ the set $\mathcal{A}_{c+\epsilon}^{y,y'} \setminus \mathcal{A}_c^{y,y'}$ is empty.

If $\boldsymbol{\ell}_y - \boldsymbol{\ell}_{y'}$ is not a scalar multiple of $\mathbf{e}$, then $\mathbf{p}^\top (\boldsymbol{\ell}_y - \boldsymbol{\ell}_{y'})$ is a non-constant linear function of $\mathbf{p}$ over $\Delta_n$. From Lemma 22, $\lambda\left(\mathcal{A}_{c+\epsilon}^{y,y'} \setminus \mathcal{A}_c^{y,y'}\right)$ goes to zero as $\epsilon$ goes to zero. And by the absolute continuity of $\nu$ w.r.t. $\lambda$, we have $\nu\left(\mathcal{A}_{c+\epsilon}^{y,y'} \setminus \mathcal{A}_c^{y,y'}\right)$ goes to zero as $\epsilon$ goes to zero.

As the above arguments hold for any $c \in [0, C]$ and $y, y' \in [n]$ with $y \neq y'$, the proof of part (b) is complete.

*Part (c):*

We have,

$$\mathcal{A}_0 \subseteq \bigcup_{y,y' \in [n], y \neq y'} \left( \mathcal{A}_0^{y,y'} \cap \mathcal{A}_0^{y',y} \right).$$

To show $r(0) = 0$, we show $\lambda\left( \mathcal{A}_0^{y,y'} \cap \mathcal{A}_0^{y',y} \right) = 0$ for all $y \neq y'$. Let $y, y' \in [n]$ with $y \neq y'$, then

$$\left( \mathcal{A}_0^{y,y'} \cap \mathcal{A}_0^{y',y} \right) = \{ \mathbf{p} \in \Delta_n : \mathbf{p}^\top (\boldsymbol{\ell}_y - \boldsymbol{\ell}_{y'}) = 0 \}.$$

If $\boldsymbol{\ell}_y - \boldsymbol{\ell}_{y'} = d\mathbf{e}$ for some $d \neq 0$, the above set is clearly empty. If $\boldsymbol{\ell}_y - \boldsymbol{\ell}_{y'}$ is not a scalar multiple of $\mathbf{e}$, then $\mathbf{p}^\top (\boldsymbol{\ell}_y - \boldsymbol{\ell}_{y'})$ is a non-constant linear function of $\mathbf{p}$ over $\Delta_n$, and hence by Lemma 22, we have that $\lambda\left( \mathcal{A}_0^{y,y'} \cap \mathcal{A}_0^{y',y} \right) = 0$. By the absolute continuity of $\nu$ w.r.t. $\lambda$ we have that $\nu\left( \mathcal{A}_0^{y,y'} \cap \mathcal{A}_0^{y',y} \right) = 0$.

As the above arguments hold for any $y, y' \in [n]$ with $y \neq y'$, the proof of part (c) is complete. □

**Lemma 24 (The uniqueness of $\psi^{\mathbf{L}}$-optimal classifier).** *Let the distribution $D$ over $\mathcal{X} \times \mathcal{Y}$ be such that the measure over conditional probabilities $\nu$, is absolutely continuous w.r.t. the base measure $\lambda$. Let $\mathbf{L} \in \mathbb{R}^{n \times n}$ be such that no two columns are identical. Then, all $\psi^{\mathbf{L}}$ optimal classifiers have the same confusion matrix. i.e. the minimizer over $\mathcal{C}_D$ of $\langle \mathbf{L}, \mathbf{C} \rangle$ is unique.*

*Proof.* If $x \in \mathcal{X}$ is such that $\operatorname{argmin}_{y \in [n]} \boldsymbol{\eta}(x)^\top \boldsymbol{\ell}_y$ is a singleton, then any $\psi^{\mathbf{L}}$-optimal classifier $h^*$ is such that

$$h^*(x) = \operatorname{argmin}_{y \in [n]} \boldsymbol{\eta}(x)^\top \boldsymbol{\ell}_y$$

We just show that the set of instances in $\mathcal{X}$, such that $\operatorname{argmin}_{y \in [n]} \boldsymbol{\eta}(x)^\top \boldsymbol{\ell}_y$ is not a singleton, has measure zero. For any $\mathbf{v} \in \mathbb{R}^n$, let $(\mathbf{v})_{(i)}$ be the $i^{th}$ element when the components of $\mathbf{v}$ are arranged in ascending order.

$$\begin{aligned} \mu\left( \{ x \in \mathcal{X} : |\operatorname{argmin}_{y \in [n]} \boldsymbol{\eta}(x)^\top \boldsymbol{\ell}_y| > 1 \} \right) &= \mu\left( \{ x \in \mathcal{X} : (\boldsymbol{\eta}(x)^\top \mathbf{L})_{(1)} = (\boldsymbol{\eta}(x)^\top \mathbf{L})_{(2)} \} \right) \\ &= \nu\left( \{ \mathbf{p} \in \Delta_n : (\mathbf{p}^\top \mathbf{L})_{(1)} = (\mathbf{p}^\top \mathbf{L})_{(2)} \} \right) \end{aligned}$$

Thus, by Lemma 23 (part c), we have that set of instances in $\mathcal{X}$ such that $\operatorname{argmin}_{y \in [n]} \boldsymbol{\eta}(x)^\top \boldsymbol{\ell}_y$ is not a singleton, has measure zero. Thus any pair of $\psi^{\mathbf{L}}$-optimal classifiers are same $\mu$ almost everywhere, and hence all the $\psi^{\mathbf{L}}$-optimal classifiers have the same confusion matrix. □

Next we give the master Lemma which uses every result in this section, and will actually be the only tool in the proofs of Lemma 12 and Theorem 13.

**Lemma 25 (Master Lemma).** *Let the distribution $D$ over $\mathcal{X} \times \mathcal{Y}$ be such that the measure over conditional probabilities $\nu$, be absolutely continuous w.r.t. the base measure $\lambda$. Let $\mathbf{L} \in [0,1]^{n \times n}$ be such that no two columns are identical. Then,*

$$\operatorname{argmin}_{\mathbf{C} \in \overline{\mathcal{C}_D}} \langle \mathbf{L}, \mathbf{C} \rangle = \operatorname{argmin}_{\mathbf{C} \in \mathcal{C}_D} \langle \mathbf{L}, \mathbf{C} \rangle.$$

*Moreover, the above set is a singleton.*

*Proof.* The first part of the proof where one shows $\operatorname{argmin}_{\mathbf{C} \in \mathcal{C}_D} \langle \mathbf{L}, \mathbf{C} \rangle$ is a singleton is exactly what is given by Lemma 24. Let

$$\mathbf{C}^* = \mathbf{C}^D[h^*] = \operatorname{argmin}_{\mathbf{C} \in \mathcal{C}_D} \langle \mathbf{L}, \mathbf{C} \rangle.$$

The classifier $h^* : \mathcal{X} \to \Delta_n$ is such that $\mathbf{C}^D[h^*] = \mathbf{C}^*$, and is fixed for convenience as the following classifier,

$$h^*(x) = \text{rand}(\operatorname{argmin}^*_{y \in [n]} \boldsymbol{\eta}(x)^\top \boldsymbol{\ell}_y).$$

Also let $\mathbf{f}^* : \Delta_n \to \Delta_n$ be such that, $h^* = \mathbf{f}^* \circ \boldsymbol{\eta}$, i.e.

$$\mathbf{f}^*(\mathbf{p}) = \text{rand}(\text{argmin}^*_{y \in [n]} \, \mathbf{p}^\top \boldsymbol{\ell}_y) \; .$$

In the rest of the proof we simply show that $\mathbf{C}^*$ is the unique minimizer of $\langle \mathbf{L}, \mathbf{C} \rangle$ over $\overline{\mathcal{C}_D}$ as well. To do so, we first assume that $\mathbf{C}' \in \text{argmin}_{\mathbf{C} \in \overline{\mathcal{C}_D}} \langle \mathbf{L}, \mathbf{C} \rangle$ and $\mathbf{C}' \neq \mathbf{C}^*$. We then go on to show a contradiction, the brief details of which are given below:

1. As $\mathbf{C}' \neq \mathbf{C}^*$, we have $||\mathbf{C}^* - \mathbf{C}'||_1 = \xi > 0$.
2. As $\mathbf{C}' \in \overline{\mathcal{C}_D}$, there exists a sequence of classifiers $h_1, h_2, \ldots$, such that their confusion matrices converge to $\mathbf{C}'$.
3. The confusion matrices of the classifiers $h_1, h_2, \ldots$, are bounded away from $\mathbf{C}^*$ as $\xi > 0$.
4. Due to the continuity of the $\mathbf{C}^D$ mapping (Lemma 21), the classifiers $h_1, h_2, \ldots$ are also bounded away from $h^*$ – i.e. they must predict differently from $h^*$ on a significant fraction of the instances.
5. Due to Lemma 23, we have that for most instances the second best prediction (in terms of loss $\mathbf{L}$) is significantly worse than the best prediction. The classifiers $h_1, h_2, \ldots$ all predict differently from $h^*$ (which always predicts the best label for any given instance) for a large fraction of the instances, hence they must predict a significantly worse label for a large fraction of instances.
6. From the above reasoning, the classifiers $h_1, h_2, \ldots$, all perform worse by a constant additive factor than $h^*$, on the $\psi^{\mathbf{L}}$ performance measure. But, as the the confusion matrices of these classifiers converge to $\mathbf{C}'$, the $\psi^{\mathbf{L}}$ performance of these classifiers must approach the optimal. Thus providing a contradiction.

The full details of the above sketch is given below.

Let

$$||\mathbf{C}' - \mathbf{C}^*||_1 = \xi > 0 \; .$$

As $\mathbf{C}' \in \overline{\mathcal{C}_D}$, we have that for all $\epsilon > 0$, there exists $\mathbf{C}_\epsilon \in \mathcal{C}_D$, such that $||\mathbf{C}_\epsilon - \mathbf{C}'||_1 \leq \epsilon$. By triangle inequality, this implies that

$$||\mathbf{C}_\epsilon - \mathbf{C}^*||_1 \quad \geq \quad \xi - \epsilon \; , \tag{3}$$

Let $\mathbf{f}_\epsilon : \Delta_n \to \Delta_n$ be s.t. $\mathbf{C}_\epsilon = \mathbf{C}^D[\mathbf{f}_\epsilon \circ \boldsymbol{\eta}]$. Now we describe the set of conditional probabilities $\mathbf{p} \in \Delta_n$ for which $\mathbf{f}_\epsilon(\mathbf{p})$ differs significantly from $\mathbf{f}^*(\mathbf{p})$. Denote this 'bad' set as $\mathcal{B}$. Let

$$\mathcal{B} = \{\mathbf{p} \in \Delta_n : ||\mathbf{f}^*(\mathbf{p}) - \mathbf{f}_\epsilon(\mathbf{p})||_1 \geq \frac{\xi}{4}\} \; .$$

We now show that this set is 'large'. Applying Eq. 3 and Lemma 21 we have

$$\begin{aligned}
\xi - \epsilon &\leq \left|\left|\mathbf{C}^D[\mathbf{f}^* \circ \boldsymbol{\eta}] - \mathbf{C}^D[\mathbf{f}_\epsilon \circ \boldsymbol{\eta}]\right|\right|_1 \\
&\leq \int_{\mathbf{p} \in \Delta_n} ||\mathbf{f}^*(\mathbf{p}) - \mathbf{f}_\epsilon(\mathbf{p})||_1 d\nu(\mathbf{p}) \\
&\leq \int_{\mathbf{p} \in \mathcal{B}} 2 d\nu(\mathbf{p}) + \int_{\mathbf{p} \notin \mathcal{B}} \frac{\xi}{4} d\nu(\mathbf{p}) \\
&= 2\nu(\mathcal{B}) + \frac{\xi}{4}(1 - \nu(\mathcal{B})) \\
&\leq 2\nu(\mathcal{B}) + \frac{\xi}{4} \\
\nu(\mathcal{B}) &\geq \frac{3\xi}{8} - \frac{\epsilon}{2}
\end{aligned} \tag{4}$$

Now we consider the set of conditional probabilities $\mathbf{p} \in \Delta_n$, such that the second best and best predictions (in terms of $\mathbf{L}$) are 'close' in performance. We show that this set is 'small'.

For any $c > 0$, define $\mathcal{A}_c \subseteq \Delta_n$ as

$$\mathcal{A}_c = \{\mathbf{p} \in \Delta_n : (\mathbf{p}^\top \mathbf{L})_{(2)} - (\mathbf{p}^\top \mathbf{L})_{(1)} \leq c\} \, .$$

From Lemma 23 we have that $\nu(\mathcal{A}_c)$ is a continuous function of $c$ close to 0 and $\nu(\mathcal{A}_0) = 0$. Let $c > 0$ be such that

$$\nu(\mathcal{A}_c) \leq \frac{\xi}{16} \, . \tag{5}$$

From Eq. 4 and 5, we have

$$\nu(\mathcal{B} \setminus \mathcal{A}_c) \geq \frac{5\xi}{16} - \frac{\epsilon}{2}$$

Any $\mathbf{p} \in \mathcal{B} \setminus \mathcal{A}_c$ is such that $\mathbf{f}_\epsilon(\mathbf{p})$ is different from $\mathbf{f}^*(\mathbf{p})$, and the second best prediction is significantly worse than the best prediction, i.e.

$$(\mathbf{p}^\top \mathbf{L})_{(2)} - (\mathbf{p}^\top \mathbf{L})_{(1)} > c \quad \text{and} \quad ||\mathbf{f}^*(\mathbf{p}) - \mathbf{f}_\epsilon(\mathbf{p})||_1 \geq \frac{\xi}{4} \, .$$

For any $\mathbf{p} \in \Delta_n$, we have $\mathbf{f}^*(\mathbf{p}) \in \Delta_n$, has a 1 at the index $\operatorname{argmin}^*_{y \in [n]} \mathbf{p}^\top \boldsymbol{\ell}_y$ and zero elsewhere. For any $\mathbf{p} \in \mathcal{B} \setminus \mathcal{A}_c$, we have $||\mathbf{f}^*(\mathbf{p}) - \mathbf{f}_\epsilon(\mathbf{p})||_1 \geq \frac{\xi}{4}$, and hence the value of $\mathbf{f}_\epsilon(\mathbf{p})$ corresponding to the index $\operatorname{argmin}^*_{y \in [n]} \mathbf{p}^\top \boldsymbol{\ell}_y$, is at most $(1 - \frac{\xi}{8})$. In particular, we have

$$\mathbf{p}^\top \mathbf{L} \mathbf{f}_\epsilon(\mathbf{p}) \geq \left(1 - \frac{\xi}{8}\right) (\mathbf{p}^\top \mathbf{L})_{(1)} + \left(\frac{\xi}{8}\right) (\mathbf{p}^\top \mathbf{L})_{(2)} \, . \tag{6}$$

By using Lemma 19 and Eq. 6 we have,

$$\begin{aligned}
\langle \mathbf{L}, \mathbf{C}_\epsilon \rangle - \langle \mathbf{L}, \mathbf{C}^* \rangle &= \int_{\mathbf{p} \in \Delta_n} \mathbf{p}^\top \mathbf{L} (\mathbf{f}_\epsilon(\mathbf{p}) - \mathbf{f}^*(\mathbf{p})) d\nu(\mathbf{p}) \\
&= \int_{\mathbf{p} \in \mathcal{B} \setminus \mathcal{A}_c} \mathbf{p}^\top \mathbf{L} (\mathbf{f}_\epsilon(\mathbf{p}) - \mathbf{f}^*(\mathbf{p})) d\nu(\mathbf{p}) + \int_{\mathbf{p} \in \Delta_n \setminus (\mathcal{B} \setminus \mathcal{A}_c)} \mathbf{p}^\top \mathbf{L} (\mathbf{f}_\epsilon(\mathbf{p}) - \mathbf{f}^*(\mathbf{p})) d\nu(\mathbf{p}) \\
&\geq \int_{\mathbf{p} \in \mathcal{B} \setminus \mathcal{A}_c} \mathbf{p}^\top \mathbf{L} (\mathbf{f}_\epsilon(\mathbf{p}) - \mathbf{f}^*(\mathbf{p})) d\nu(\mathbf{p}) \\
&\geq \int_{\mathbf{p} \in \mathcal{B} \setminus \mathcal{A}_c} \left( \left(1 - \frac{\xi}{8}\right) (\mathbf{p}^\top \mathbf{L})_{(1)} + \left(\frac{\xi}{8}\right) (\mathbf{p}^\top \mathbf{L})_{(2)} - (\mathbf{p}^\top \mathbf{L})_{(1)} \right) d\nu(\mathbf{p}) \\
&= \int_{\mathbf{p} \in \mathcal{B} \setminus \mathcal{A}_c} \frac{\xi}{8} \left( (\mathbf{p}^\top \mathbf{L})_{(2)} - (\mathbf{p}^\top \mathbf{L})_{(1)} \right) d\nu(\mathbf{p}) \\
&\geq \frac{\xi c}{8} \left( \frac{5\xi}{16} - \frac{\epsilon}{2} \right)
\end{aligned}$$

If $\epsilon \leq \frac{\xi}{2}$, we have

$$\langle \mathbf{L}, \mathbf{C}_\epsilon \rangle - \langle \mathbf{L}, \mathbf{C}^* \rangle \geq \frac{\xi^2 c}{128} \, . \tag{7}$$

The above holds for any $\epsilon \in (0, \frac{\xi}{2}]$, and both $\xi$ and $c$ do not depend on $\epsilon$.

However, we have $\mathbf{C}' \in \operatorname{argmin}_{\mathbf{C} \in \overline{\mathcal{C}_D}} \langle \mathbf{L}, \mathbf{C} \rangle$, and $||\mathbf{C}_\epsilon - \mathbf{C}'||_1 \leq \epsilon$. Hence,

$$\begin{aligned}
\langle \mathbf{L}, \mathbf{C}_\epsilon \rangle &= \langle \mathbf{L}, \mathbf{C}' \rangle + \langle \mathbf{L}, \mathbf{C}_\epsilon - \mathbf{C}' \rangle \\
&\leq \langle \mathbf{L}, \mathbf{C}' \rangle + ||\mathbf{L}||_\infty ||\mathbf{C}_\epsilon - \mathbf{C}'||_1
\end{aligned}$$

$$\begin{aligned} &\leq \langle \mathbf{L}, \mathbf{C}' \rangle + \epsilon \\ &\leq \min_{\mathbf{C} \in \mathcal{C}_D} \langle \mathbf{L}, \mathbf{C} \rangle + \epsilon \\ &= \langle \mathbf{L}, \mathbf{C}^* \rangle + \epsilon \end{aligned} \tag{8}$$

It can be clearly seen that, for small enough $\epsilon$, Eqs. 7 and 8 contradict each other. And thus we have that $\mathbf{C}' = \mathbf{C}^*$. □

### B.5. Proof of Lemma 12

**Lemma** (Existence of Bayes optimal classifier for monotonic $\psi$)**.** *Let $D$ be such that the probability measure associated with the random vector $\boldsymbol{\eta}(X) = (\eta_1(X), \ldots, \eta_n(X))^\top$ is absolutely continuous w.r.t. the base probability measure associated with the uniform distribution over $\Delta_n$, and let $\psi$ be a performance measure that is differentiable and bounded over $\mathcal{C}_D$, and is monotonically increasing in $C_{ii}$ for each $i$ and non-increasing in $C_{ij}$ for all $i, j$. Then $\exists h^* : \mathcal{X} \to \Delta_n$ s.t. $\mathcal{P}_D^\psi[h^*] = \mathcal{P}_D^{\psi,*}$.*

*Proof.* Let $\mathbf{C}^* = \operatorname{argmax}_{\mathbf{C} \in \overline{\mathcal{C}_D}} \psi(\mathbf{C})$. Such a $\mathbf{C}^*$ always exists by compactness of $\overline{\mathcal{C}_D}$ and continuity of $\psi$. We will show that this $\mathbf{C}^*$ is also in $\mathcal{C}_D$, thus proving the existence of $h^* : \mathcal{X} \to \Delta_n$ which is such that $\mathbf{C}^* = \mathbf{C}^D[h^*]$ and hence $\mathcal{P}_D^\psi[h^*] = \mathcal{P}_D^{\psi,*}$.

By first order optimality, and convexity of $\overline{\mathcal{C}_D}$, we have that for all $\mathbf{C} \in \overline{\mathcal{C}_D}$

$$\langle \nabla \psi(\mathbf{C}^*), \mathbf{C}^* \rangle \geq \langle \nabla \psi(\mathbf{C}^*), \mathbf{C} \rangle \ .$$

Let $\mathbf{L}^*$ be the scaled and shifted version of $-\nabla \psi(\mathbf{C}^*)$ with entries in $[0, 1]$, then we have that

$$\mathbf{C}^* \in \operatorname{argmin}_{\mathbf{C} \in \overline{\mathcal{C}_D}} \langle \mathbf{L}^*, \mathbf{C} \rangle \ .$$

Due to the monotonicity condition on $\psi$ the diagonal elements of its gradient $\nabla \psi(\mathbf{C}^*)$ are positive, and the off-diagonal elements are non-positive, and hence no two columns of $\mathbf{L}^*$ are identical. Thus by Lemma 25, we have that $\mathbf{C}^* \in \mathcal{C}_D$. □

### B.6. Proof of Theorem 13

**Theorem** (Form of Bayes optimal classifier for monotonic $\psi$)**.** *Let $D$, $\psi$ satisfy the conditions of Lemma 12. Let $h^* : \mathcal{X} \to \Delta_n$ be a $\psi$-optimal classifier and let $\mathbf{C}^* = \mathbf{C}^D[h^*]$. Let $\widetilde{\mathbf{L}}^* = -\nabla \psi(\mathbf{C}^*)$, and let $\mathbf{L}^* \in [0, 1]^{n \times n}$ be obtained by scaling and shifting $\widetilde{\mathbf{L}}^*$ so its entries lie in $[0, 1]$. Then any classifier that is $\psi^{\mathbf{L}^*}$-optimal is also $\psi$-optimal.*

*Proof.* Clearly

$$\psi(\mathbf{C}^*) = \max_{\mathbf{C} \in \mathcal{C}_D} \psi(\mathbf{C}) \ .$$

Hence by the differentiability of $\psi$, first order conditions for optimality and convexity of $\mathcal{C}_D$ we have $\forall \mathbf{C} \in \mathcal{C}_D$,

$$\langle \nabla \psi(\mathbf{C}^*), \mathbf{C}^* \rangle \geq \langle \nabla \psi(\mathbf{C}^*), \mathbf{C} \rangle \ .$$

By definition of $\mathbf{L}^*$, this implies that $\forall \mathbf{C} \in \mathcal{C}_D$,

$$\langle \mathbf{L}^*, \mathbf{C}^* \rangle \leq \langle \mathbf{L}^*, \mathbf{C} \rangle \ .$$

Thus, we have that $h^*$ is a $\psi^{\mathbf{L}^*}$-optimal classifier.

Due to the monotonicity condition on $\psi$ the diagonal elements of its gradient $\nabla \psi(\mathbf{C}^*)$ are positive, and the off-diagonal elements are non-positive, and hence no two columns of $\mathbf{L}^*$ are identical. By Lemma 24 (or Lemma 25), we have that all $\psi^{\mathbf{L}^*}$-optimal classifiers have the same confusion matrix, which is equal to $\mathbf{C}^D[h^*] = \mathbf{C}^*$. And thus all $\psi^{\mathbf{L}^*}$ optimal classifiers are also $\psi$-optimal. □

# C. Supplementary Material for Section 5 (Consistency)

## C.1. Proof of Lemma 14

**Lemma** (**L**-regret of multiclass plug-in classifiers)**.** *For a fixed* $\mathbf{L} \in [0,1]^{n\times n}$ *and class probability estimation model* $\widehat{\boldsymbol{\eta}} : \mathcal{X}\to\Delta_n$*, let* $\widehat{h} : \mathcal{X}\to[n]$ *be a classifier* $\widehat{h}(x) = \operatorname{argmin}^*_{j\in[n]} \sum_{i=1}^n \widehat{\eta}_i(x) L_{ij}$*. Then*

$$\mathcal{P}_D^{\mathbf{L},*} - \mathcal{P}_D^{\mathbf{L}}[\widehat{h}] \leq \mathbf{E}_X\left[\left\|\widehat{\boldsymbol{\eta}}(X) - \boldsymbol{\eta}(X)\right\|_1\right].$$

*Proof.* Let $h^* : \mathcal{X}\to\Delta_n$ be such that

$$h^*(x) = \operatorname{argmin}^*_{y\in[n]} \boldsymbol{\ell}_y^\top \boldsymbol{\eta}(x) \,.$$

By Proposition 6, we have that

$$h^* \in \operatorname{argmax}_{h:\mathcal{X}\to\Delta_n} \mathcal{P}_D^{\mathbf{L}}[h] \,.$$

We then have

$$\begin{aligned}
\mathcal{P}_D^{\mathbf{L},*} - \mathcal{P}_D^{\mathbf{L}}[\widehat{h}] &= \langle \mathbf{L}, \mathbf{C}^D[\widehat{h}]\rangle - \langle \mathbf{L}, \mathbf{C}^D[h^*]\rangle \\
&= \mathbf{E}_X\left[\boldsymbol{\eta}(X)^\top \boldsymbol{\ell}_{\widehat{h}(X)}\right] - \mathbf{E}_X\left[\boldsymbol{\eta}(X)^\top \boldsymbol{\ell}_{h^*(X)}\right] \\
&= \mathbf{E}_X\left[\widehat{\boldsymbol{\eta}}(X)^\top \boldsymbol{\ell}_{\widehat{h}(X)}\right] + \mathbf{E}_X\left[(\boldsymbol{\eta}(X) - \widehat{\boldsymbol{\eta}}(X)^\top \boldsymbol{\ell}_{\widehat{h}(X)}\right] - \mathbf{E}_X\left[\boldsymbol{\eta}(X)^\top \boldsymbol{\ell}_{h^*(X)}\right] \\
&\leq \mathbf{E}_X\left[\widehat{\boldsymbol{\eta}}(X)^\top \boldsymbol{\ell}_{h^*(X)}\right] + \mathbf{E}_X\left[(\boldsymbol{\eta}(X) - \widehat{\boldsymbol{\eta}}(X)^\top \boldsymbol{\ell}_{\widehat{h}(X)}\right] - \mathbf{E}_X\left[\boldsymbol{\eta}(X)^\top \boldsymbol{\ell}_{h^*(X)}\right] \\
&= \mathbf{E}_X\left[(\boldsymbol{\eta}(X) - \widehat{\boldsymbol{\eta}}(X))^\top (\boldsymbol{\ell}_{\widehat{h}(X)} - \boldsymbol{\ell}_{h^*(X)})\right] \\
&\leq \mathbf{E}_X\left[\left\|\boldsymbol{\eta}(X) - \widehat{\boldsymbol{\eta}}(X)\right\|_1 \cdot \left\|\boldsymbol{\ell}_{\widehat{h}(X)} - \boldsymbol{\ell}_{h^*(X)}\right\|_\infty\right] \\
&\leq \mathbf{E}_X\left[\left\|\boldsymbol{\eta}(X) - \widehat{\boldsymbol{\eta}}(X)\right\|_1\right],
\end{aligned}$$

as desired. □

## C.2. Proof of Lemma 15

**Lemma** (Uniform convergence of confusion matrices)**.** *For* $\mathbf{q} : \mathcal{X}\to\Delta_n$*, let*

$$\mathcal{H}_{\mathbf{q}} = \{h:\mathcal{X}\to[n],\, h(x) = \operatorname{argmin}^*_{j\in[n]} \sum_{i=1}^n q_i(x) L_{ij} \,|\, \mathbf{L} \in [0,1]^{n\times n}\} \,.$$

*Let* $S \in (\mathcal{X}\times[n])^m$ *be a sample drawn i.i.d. from* $D^m$*. For any* $\delta \in (0,1]$*, w.p. at least* $1-\delta$ *(over draw of* $S$ *from* $D^m$*),*

$$\sup_{h\,\in\,\mathcal{H}_{\mathbf{q}}} \left\|\mathbf{C}^D[h] - \widehat{\mathbf{C}}^S[h]\right\|_\infty \leq C\sqrt{\frac{n^2\log(n)\log(m) + \log(n^2/\delta)}{m}},$$

*where* $C > 0$ *is a distribution-independent constant.*

*Proof.* For any $a, b \in [n]$ we have,

$$\begin{aligned}
\sup_{h\in\mathcal{H}_{\mathbf{q}}} \left|\widehat{C}^S_{a,b}[h] - C^D_{a,b}[h]\right| &= \sup_{h\in\mathcal{H}_{\mathbf{q}}} \left|\frac{1}{m}\sum_{i=1}^m (\mathbf{1}(y_i = a, h(x_i) = b) - \mathbf{E}[\mathbf{1}(Y = a, h(X) = b)])\right| \\
&= \sup_{h\in\mathcal{H}^b_{\mathbf{q}}} \left|\frac{1}{m}\sum_{i=1}^m (\mathbf{1}(y_i = a, h(x_i) = 1) - \mathbf{E}[\mathbf{1}(Y = a, h(X) = 1)])\right| \,,
\end{aligned}$$

where for a fixed $b \in [n]$, $\mathcal{H}^b_{\mathbf{q}} = \{h : \mathcal{X}\to\{0,1\} : \exists \mathbf{L} \in [0,1]^{n\times n}, \forall x \in \mathcal{X}, h(x) = \mathbf{1}(b = \operatorname{argmin}^*_{y\in[n]} \boldsymbol{\ell}_y^\top \mathbf{q}(x))\}$. The set $\mathcal{H}^b_{\mathbf{q}}$ can be seen as hypothesis class whose concepts are the intersection of $n$ halfspaces in $\mathbb{R}^n$ (corresponding to $\mathbf{q}(x)$)

through the origin. Hence we have from Lemma 3.2.3 of Blumer et al. (1989) that the VC-dimension of $\mathcal{H}^b_{\mathbf{q}}$ is at most $2n^2\log(3n)$. From standard uniform convergence arguments we have that for each $a, b \in [n]$, the following holds with at least probability $1-\delta$,

$$\sup_{h\in\mathcal{H}_{\mathbf{q}}} \left|\widehat{C}^S_{a,b}[h] - C^D_{a,b}[h]\right| \leq C\sqrt{\frac{n^2\log(n)\log(m)+\log(\frac{1}{\delta})}{m}}$$

where $C>0$ is some constant. Applying union bound over all $a, b \in [n]$ we have that the following holds with probability at least $1-\delta$

$$\sup_{h\in\mathcal{H}_{\mathbf{q}}} \left|\left|\widehat{\mathbf{C}}^S[h] - \mathbf{C}^D[h]\right|\right|_\infty \leq C\sqrt{\frac{n^2\log(n)\log(m)+\log(\frac{n^2}{\delta})}{m}}\,.$$

□

### C.3. Proof of Theorem 16

**Theorem** ($\psi$-regret of Frank-Wolfe method based algorithm)**.** *Let $\psi : [0,1]^{n\times n}\to\mathbb{R}_+$ be concave over $\mathcal{C}_D$, and L-Lipschitz and $\beta$-smooth w.r.t. the $\ell_1$ norm. Let $S = (S_1, S_2) \in (\mathcal{X}\times[n])^m$ be a training sample drawn i.i.d. from $D$. Further, let $\widehat{\boldsymbol{\eta}} : \mathcal{X}\to\Delta_n$ be the* CPE *model learned from $S_1$ in Algorithm 1 and $h^{\text{FW}}_S : \mathcal{X}\to\Delta_n$ be the classifier obtained after $\kappa m$ iterations. Then for any $\delta\in(0,1]$, with probability at least $1-\delta$ (over draw of $S$ from $D^m$),*

$$\mathcal{P}^{\psi,*}_D - \mathcal{P}^{\psi}_D[h^{\text{FW}}_S] \leq 4L\mathbf{E}_X\left[\left\|\widehat{\boldsymbol{\eta}}(X)-\boldsymbol{\eta}(X)\right\|_1\right] + 4\sqrt{2}\beta n^2 C\sqrt{\frac{n^2\log(n)\log(m)+\log(n^2/\delta)}{m}} + \frac{8\beta}{\kappa m+2},$$

*where $C>0$ is a distribution-independent constant.*

We first prove an important lemma where we bound the approximation error of the linear optimization oracle used in the algorithm using Lemma 14 and 15. This result coupled with the standard convergence analysis for the Frank-Wolfe method (Jaggi, 2013) will then allow us to prove the above theorem.

**Lemma 26.** *Let $\psi : [0,1]^{n\times n}\to\mathbb{R}_+$ be concave over $\mathcal{C}_D$, and L-Lipschitz and $\beta$-smooth w.r.t. the $\ell_1$ norm. Let classifiers $\widehat{g}^1,\ldots,\widehat{g}^T$, and $h^0, h^1,\ldots,h^T$ be as defined in Algorithm 1. Then for any $\delta\in(0,1]$, with probability at least $1-\delta$ (over draw of $S$ from $D^m$), we have for all $1\leq t\leq T$*

$$\langle\nabla\psi(\mathbf{C}^D[h^{t-1}]), \mathbf{C}^D[\widehat{g}^t]\rangle \geq \max_{\mathbf{g}:\mathcal{X}:\to\Delta_n}\langle\nabla\psi(\mathbf{C}^D[h^{t-1}]), \mathbf{C}^D[\mathbf{g}]\rangle - \epsilon_S$$

*where*

$$\epsilon_S = 2L\mathbf{E}_X\left[\left\|\boldsymbol{\eta}(X)-\widehat{\boldsymbol{\eta}}(X)\right\|_1\right] + 2\sqrt{2}C\beta n^2\sqrt{\frac{n^2\log(n)\log(m)+\log(\frac{n^2}{\delta})}{m}}\,.$$

*Proof.* For any $1\leq t\leq T$, let $\mathbf{g}^{t,*}\in\text{argmax}_{\mathbf{g}:\mathcal{X}\to\Delta_n}\langle\nabla\psi(\mathbf{C}^D[h^{t-1}]), \mathbf{C}^D[\mathbf{g}]\rangle$. We then have

$$\begin{aligned}
&\max_{\mathbf{g}:\mathcal{X}:\to\Delta_n}\langle\nabla\psi(\mathbf{C}^D[h^{t-1}]), \mathbf{C}^D[\mathbf{g}]\rangle - \langle\nabla\psi(\mathbf{C}^D[h^{t-1}]), \mathbf{C}^D[\widehat{g}^t]\rangle\\
&= \langle\nabla\psi(\mathbf{C}^D[h^{t-1}]), \mathbf{C}^D[\mathbf{g}^{t,*}]\rangle - \langle\nabla\psi(\mathbf{C}^D[h^{t-1}]), \mathbf{C}^D[\widehat{g}^t]\rangle\\
&= \underbrace{\langle\nabla\psi(\mathbf{C}^D[h^{t-1}]), \mathbf{C}^D[\mathbf{g}^{t,*}]\rangle - \langle\nabla\psi(\widehat{\mathbf{C}}^{S_2}[h^{t-1}]), \mathbf{C}^D[g^{t,*}]\rangle}_{\text{term}_1}\\
&\qquad + \underbrace{\langle\nabla\psi(\widehat{\mathbf{C}}^{S_2}[h^{t-1}]), \mathbf{C}^D[g^{t,*}]\rangle - \langle\nabla\psi(\widehat{\mathbf{C}}^{S_2}[h^{t-1}]), \mathbf{C}^D[\widehat{g}^t]\rangle}_{\text{term}_2}\\
&\qquad + \underbrace{\langle\nabla\psi(\widehat{\mathbf{C}}^{S_2}[h^{t-1}]), \mathbf{C}^D[\widehat{g}^t]\rangle - \langle\nabla\psi(\mathbf{C}^D[h^{t-1}]), \mathbf{C}^D[\widehat{g}^t]\rangle}_{\text{term}_3}.
\end{aligned}$$

We next bound each of these terms. We start with term$_2$. For any $1 \leq t \leq T$, let $\widehat{\mathbf{L}}^t$ be as defined in Algorithm 1. Since $\widehat{\mathbf{L}}^t$ is a scaled and translated version of the gradient $\nabla\psi(\widehat{\mathbf{C}}^{S_2}[h^{t-1}])$, we have $\nabla\psi(\widehat{\mathbf{C}}^{S_2}[h^{t-1}]) = c_t - a_t\widehat{\mathbf{L}}^t$, for some constant $c_t \in \mathbb{R}$ and $a_t \in [0, 2L]$. Thus for all $1 \leq t \leq T$,

$$\begin{aligned}
&\langle\nabla\psi(\widehat{\mathbf{C}}^{S_2}[h^{t-1}]), \mathbf{C}^D[\mathbf{g}^{t,*}]\rangle - \langle\nabla\psi(\widehat{\mathbf{C}}^{S_2}[h^{t-1}]), \mathbf{C}^D[\widehat{g}^t]\rangle \\
&\qquad = \; a_t \cdot (\langle\widehat{\mathbf{L}}^t, \mathbf{C}^D[\widehat{g}^t]\rangle - \langle\widehat{\mathbf{L}}^t, \mathbf{C}^D[\mathbf{g}^{t,*}]\rangle) \\
&\qquad = \; a_t \cdot (\mathcal{P}_D^{\widehat{\mathbf{L}}^t}[\mathbf{g}^{t,*}] - \mathcal{P}_D^{\widehat{\mathbf{L}}^t}[\widehat{g}^t]) \\
&\qquad \leq \; a_t \cdot (\mathcal{P}_D^{\widehat{\mathbf{L}}^t,*} - \mathcal{P}_D^{\widehat{\mathbf{L}}^t}[\widehat{g}^t]) \\
&\qquad \leq \; 2L \cdot (\mathcal{P}_D^{\widehat{\mathbf{L}}^t,*} - \mathcal{P}_D^{\widehat{\mathbf{L}}^t}[\widehat{g}^t]) \\
&\qquad \leq \; 2L \cdot \mathbf{E}_X\left[\left\|\boldsymbol{\eta}(X) - \widehat{\boldsymbol{\eta}}(X)\right\|_1\right],
\end{aligned}$$

where the third step uses the definition of $\mathcal{P}_D^{\widehat{\mathbf{L}}^t,*}$ and the last step follows from Lemma 14.

Next, for term$_1$, we have by an application of Holder's inequality

$$\begin{aligned}
&\langle\nabla\psi(\mathbf{C}^D[h^{t-1}]), \mathbf{C}^D[\mathbf{g}^{t,*}]\rangle \; - \; \langle\nabla\psi(\widehat{\mathbf{C}}^{S_2}[h^{t-1}]), \mathbf{C}^D[g^{t,*}]\rangle \\
&\qquad \leq \; \left\|\nabla\psi(\widehat{\mathbf{C}}^{S_2}[h^{t-1}]) - \nabla\psi(\mathbf{C}^D[h^{t-1}])\right\|_\infty \|\mathbf{C}^D[\mathbf{g}^{t,*}]\|_1 \\
&\qquad = \; \left\|\nabla\psi(\widehat{\mathbf{C}}^{S_2}[h^{t-1}]) - \nabla\psi(\mathbf{C}^D[h^{t-1}])\right\|_\infty (1) \\
&\qquad \leq \; \beta\left\|\widehat{\mathbf{C}}^{S_2}[h^{t-1}] - \mathbf{C}^D[h^{t-1}]\right\|_1 \\
&\qquad \leq \; \beta n^2\left\|\widehat{\mathbf{C}}^{S_2}[h^{t-1}] - \mathbf{C}^D[h^{t-1}]\right\|_\infty \\
&\qquad \leq \; \beta n^2 \max_{i\in[t-1]}\left\|\widehat{\mathbf{C}}^{S_2}[\widehat{g}^i] - \mathbf{C}^D[\widehat{g}^i]\right\|_\infty \\
&\qquad \leq \; \beta n^2 \sup_{h\in\mathcal{H}_{\widehat{\boldsymbol{\eta}}}}\left\|\widehat{\mathbf{C}}^{S_2}[h] - \mathbf{C}^D[h]\right\|_\infty,
\end{aligned}$$

where the third step follows from $\beta$-smoothness of $\psi$. One can similarly bound term$_3$. We thus have for all $1 \leq t \leq T$,

$$\begin{aligned}
&\max_{\mathbf{g}:\mathcal{X}\to\Delta_n} \langle\nabla\psi(\mathbf{C}^D[h^{t-1}]), \mathbf{C}^D[\mathbf{g}]\rangle - \langle\nabla\psi(\mathbf{C}^D[h^{t-1}]), \mathbf{C}^D[\widehat{g}^t]\rangle \\
&\qquad\qquad \leq \; 2L \cdot \mathbf{E}_X\left[\left\|\boldsymbol{\eta}(X) - \widehat{\boldsymbol{\eta}}(X)\right\|_1\right] + 2\beta n^2 \sup_{h\in\mathcal{H}_{\widehat{\boldsymbol{\eta}}}}\left\|\widehat{\mathbf{C}}^{S_2}[h] - \mathbf{C}^D[h]\right\|_\infty.
\end{aligned}$$

Applying Lemma 15 with $|S_2| = \lceil m/2\rceil$ examples, we have with probability $1-\delta$ (over random draw of $S_2$ from $D$), for all $1 \leq t \leq T$,

$$\begin{aligned}
&\max_{\mathbf{g}:\mathcal{X}\to\Delta_n} \langle\nabla\psi(\mathbf{C}^D[h^{t-1}]), \mathbf{C}^D[\mathbf{g}]\rangle - \langle\nabla\psi(\mathbf{C}^D[h^{t-1}]), \mathbf{C}^D[\widehat{g}^t]\rangle \\
&\qquad\qquad \leq \; 2L\mathbf{E}_X\left[\left\|\boldsymbol{\eta}(X) - \widehat{\boldsymbol{\eta}}(X)\right\|_1\right] + 2\sqrt{2}C\beta n^2\sqrt{\frac{n^2\log(n)\log(m) + \log(\frac{n^2}{\delta})}{m}}.
\end{aligned}$$

□

We are now ready to prove Theorem 16.

*Proof of Theorem 16.* Our proof shall make use of Lemma 26 and the standard convergence result for the Frank-Wolfe algorithm for maximizing a concave function over a convex set (Jaggi, 2013). We will find it useful to first define the following quantity, referred to as the curvature constant in (Jaggi, 2013).

$$C_\psi \;=\; \sup_{\mathbf{C}_1,\mathbf{C}_2\in\mathcal{C}_D,\gamma\in[0,1]} \frac{2}{\gamma^2}\Big(\psi\big(\mathbf{C}_1 + \gamma(\mathbf{C}_2 - \mathbf{C}_1)\big) - \psi\big(\mathbf{C}_1\big) - \gamma\big\langle\mathbf{C}_2 - \mathbf{C}_1, \nabla\psi(\mathbf{C}_1)\big\rangle\Big).$$

Also, define two positive scalars $\epsilon_S$ and $\delta_{\text{apx}}$ required in the analysis of (Jaggi, 2013):

$$\epsilon_S \;=\; 2L\mathbf{E}_X\left[\left\|\boldsymbol{\eta}(X) - \widehat{\boldsymbol{\eta}}(X)\right\|_1\right] + 2\sqrt{2}C\beta n^2\sqrt{\frac{n^2\log(n)\log(m) + \log(\frac{n^2}{\delta})}{m}}$$

$$\delta_{\text{apx}} \quad = \quad \frac{(T+1)\epsilon_S}{C_\psi},$$

where $\delta \in (0,1]$ is as in the theorem statement. Further, let the classifiers $\widehat{g}^1, \ldots, \widehat{g}^T$, and $h^0, h^1, \ldots, h^T$ be as defined in Algorithm 1. We then have from Lemma 26 that the following holds with probability at least $1-\delta$, for all $1 \leq t \leq T$,

$$\begin{aligned}
\left\langle \nabla\psi\left(\mathbf{C}^D\left[h^{t-1}\right]\right), \mathbf{C}^D\left[\widehat{g}^t\right]\right\rangle &\geq \max_{\mathbf{g}:\mathcal{X}\to\Delta_n} \left\langle \nabla\psi\left(\mathbf{C}^D\left[h^{t-1}\right]\right), \mathbf{C}^D\left[\mathbf{g}\right]\right\rangle - \epsilon_S \\
&= \max_{\mathbf{C}\in\mathcal{C}_D} \left\langle \nabla\psi\left(\mathbf{C}^D\left[h^{t-1}\right]\right), \mathbf{C}\right\rangle - \epsilon_S \\
&= \max_{\mathbf{C}\in\overline{\mathcal{C}_D}} \left\langle \nabla\psi\left(\mathbf{C}^D\left[h^{t-1}\right]\right), \mathbf{C}\right\rangle - \epsilon_S \\
&= \max_{\mathbf{C}\in\overline{\mathcal{C}_D}} \left\langle \nabla\psi\left(\mathbf{C}^D\left[h^{t-1}\right]\right), \mathbf{C}\right\rangle - \frac{1}{2}\delta_{\text{apx}}\frac{2}{T+1}C_\psi \\
&\geq \max_{\mathbf{C}\in\overline{\mathcal{C}_D}} \langle \nabla\psi\left(\mathbf{C}^D\left[h^{t-1}\right]\right), \mathbf{C}\rangle - \frac{1}{2}\delta_{\text{apx}}\frac{2}{t+1}C_\psi \ . \qquad (9)
\end{aligned}$$

Also observe that for the two sequences of iterates given by the confusion matrices of the above classifiers,

$$\mathbf{C}^D[h^t] \quad = \quad \left(1 - \frac{2}{t+1}\right)\mathbf{C}^D[h^{t-1}] + \frac{2}{t+1}\mathbf{C}^D[\widehat{g}^t], \ . \qquad (10)$$

for all $1 \leq t \leq T$. Based on Eq. (9) and Eq. (10), one can now apply the result of (Jaggi, 2013).

In particular, the sequence of iterates $\mathbf{C}^D[h^0], \mathbf{C}^D[h^1], \ldots, \mathbf{C}^D[h^T]$ can be considered as the sequence of iterates arising from running the Frank-Wolfe optimization method to maximize $\psi$ over $\overline{\mathcal{C}}_D$ with a linear optimization oracle that is $\frac{1}{2}\delta_{\text{apx}}\frac{2}{t+1}C_\psi$ accurate at iteration $t$. Since $\psi$ is a concave function over the convex constraint set $\overline{\mathcal{C}_D}$, one has from Theorem 1 in (Jaggi, 2013) that the following convergence guarantee holds with probability at least $1-\delta$:

$$\begin{aligned}
\mathcal{P}_D^\psi[h_S^{\text{FW}}] &= \psi(\mathbf{C}^D[h_S^{\text{FW}}]) \\
&= \psi(\mathbf{C}^D[h^T]) \\
&\geq \max_{\mathbf{C}\in\overline{\mathcal{C}_D}} \psi(\mathbf{C}) - \frac{2C_\psi}{T+2}(1+\delta_{\text{apx}}) \\
&= \max_{\mathbf{C}\in\overline{\mathcal{C}_D}} \psi(\mathbf{C}) - \frac{2C_\psi}{T+2} - \frac{2\epsilon_S(T+1)}{T+2} \\
&\geq \max_{\mathbf{C}\in\overline{\mathcal{C}_D}} \psi(\mathbf{C}) - \frac{2C_\psi}{T+2} - 2\epsilon_S \qquad (11)
\end{aligned}$$

We can further upper bound $C_\psi$ in the above inequality in terms of the the smoothness parameter of $\psi$:

$$\begin{aligned}
C_\psi &= \sup_{\mathbf{C}_1,\mathbf{C}_2\in\mathcal{C}_D,\gamma\in[0,1]} \frac{2}{\gamma^2}\Big(\psi\big(\mathbf{C}_1+\gamma(\mathbf{C}_2-\mathbf{C}_1)\big) - \psi\big(\mathbf{C}_1\big) - \gamma\big\langle \mathbf{C}_2-\mathbf{C}_1, \nabla\psi(\mathbf{C}_1)\big\rangle\Big) \\
&\leq \sup_{\mathbf{C}_1,\mathbf{C}_2\in\mathcal{C}_D,\gamma\in[0,1]} \frac{2}{\gamma^2}\Big(\frac{\beta}{2}\gamma^2||\mathbf{C}_1-\mathbf{C}_2||_1^2\Big) \\
&= 4\beta \ ,
\end{aligned}$$

where the second step follows from the $\beta$-smoothness of $\psi$. Substituting back in Eq. (11), we finally have with probability at least $1-\delta$,

$$\begin{aligned}
\mathcal{P}_D^\psi[h_S^{\text{FW}}] &\geq \max_{\mathbf{C}\in\overline{\mathcal{C}_D}} \psi(\mathbf{C}) - \frac{8\beta}{T+2} - 2\epsilon_S \\
&= \mathcal{P}_D^{\psi,*} - 4L\mathbf{E}_X\left[\left\|\boldsymbol{\eta}(X) - \widehat{\boldsymbol{\eta}}(X)\right\|_1\right] - 4\sqrt{2}C\beta n^2\sqrt{\frac{n^2\log(n)\log(m) + \log(\frac{n^2}{\delta})}{m}} - \frac{8\beta}{T+2},
\end{aligned}$$

which follows from the definition of $\epsilon_S$. Setting $T = \kappa m$ completes the proof. □

### C.4. Proof of Theorem 17

**Theorem.** *Let $\psi : [0,1]^{n\times n}\to\mathbb{R}_+$ be such that $\psi(\mathbf{C}) = \frac{\langle \mathbf{A},\mathbf{C}\rangle}{\langle \mathbf{B},\mathbf{C}\rangle}$ where $\mathbf{A},\mathbf{B}\in\mathbb{R}_+^{n\times n}$, $\sup_{\mathbf{C}\in\mathcal{C}_D}\psi(\mathbf{C}) \le 1$, and $\min_{\mathbf{C}\in\mathcal{C}_D}\langle \mathbf{B},\mathbf{C}\rangle \ge b$, for some $b>0$. Let $S=(S_1,S_2)\in(\mathcal{X}\times[n])^m$ be a training sample drawn i.i.d. from $D$. Let $\widehat{\boldsymbol{\eta}}:\mathcal{X}\to\Delta_n$ be the* CPE *model learned from $S_1$ in Algorithm 2 and $h_S^{\text{BS}}:\mathcal{X}\to[n]$ be the classifier obtained after $\kappa m$ iterations. Then for any $\delta\in(0,1]$, with probability at least $1-\delta$ (over draw of $S$ from $D^m$),*

$$\mathcal{P}_D^* - \mathcal{P}_D[h_S^{\text{BS}}] \le 2\tau\mathbf{E}_X\left[\left\|\widehat{\boldsymbol{\eta}}(X)-\boldsymbol{\eta}(X)\right\|_1\right] + 2\sqrt{2}C\tau\sqrt{\frac{n^2\log(n)\log(m)+\log(n^2/\delta)}{m}} + 2^{-\kappa m},$$

*where $\tau=\frac{1}{b}(\|\mathbf{A}\|_1+\|\mathbf{B}\|_1)$ and $C>0$ is a distribution-independent constant.*

We will find it useful to state the following lemmas:

**Lemma 27 (Invariant in Algorithm 2).** *Let $\psi$ be as defined in Theorem 17. Let $\mathcal{H}_{\widehat{\boldsymbol{\eta}}} = \{h : \mathcal{X}\to[n], h(x) = \operatorname{argmin}^*_{j\in[n]}\sum_{i=1}^n \widehat{\eta}_i(x)L_{ij} \mid \mathbf{L}\in[0,1]^{n\times n}\}$. Then the following invariant is true at the end of each iteration $0\le t\le T$ of Algorithm 2:*

$$\alpha^t - \tau\bar{\epsilon} < \psi(\mathbf{C}^D[h^t]) \le \sup_{\mathbf{C}\in\mathcal{C}_D}\psi(\mathbf{C}) \le \beta^t + \tau\bar{\epsilon},$$

*where $\tau=\frac{1}{b}(\|\mathbf{A}\|_1+\|\mathbf{B}\|_1)$ and $\bar{\epsilon} = \mathbf{E}_X\left[\left\|\widehat{\boldsymbol{\eta}}(X)-\boldsymbol{\eta}(X)\right\|_1\right] + \sup_{h\in\mathcal{H}_{\widehat{\boldsymbol{\eta}}}}\left\|\mathbf{C}^D[h] - \widehat{\mathbf{C}}^{S_2}[h]\right\|_\infty$.*

*Proof.* We first have from Lemma 14, the following guarantee for the linear minimization step at each iteration $t$ of Algorithm 2:

$$\begin{aligned}\langle\widehat{\mathbf{L}}^t, \mathbf{C}^D[\widehat{g}^t]\rangle &\le \min_{\mathbf{C}\in\mathcal{C}_D}\langle\widehat{\mathbf{L}}^t,\mathbf{C}\rangle + \mathbf{E}_X\left[\left\|\widehat{\boldsymbol{\eta}}(X)-\boldsymbol{\eta}(X)\right\|_1\right]\\ &= \min_{\mathbf{C}\in\mathcal{C}_D}\langle\widehat{\mathbf{L}}^t,\mathbf{C}\rangle + \epsilon \quad \text{(say)}. \end{aligned} \tag{12}$$

Further, let us denote $\epsilon' = \sup_{h\in\mathcal{H}_{\boldsymbol{\eta}}}\left\|\mathbf{C}^D[h] - \widehat{\mathbf{C}}^{S_2}[h]\right\|_\infty$. Notice that $\bar{\epsilon}=\epsilon+\epsilon'$.

We shall now prove this lemma by mathematical induction on the iteration number $t$. For $t=0$, the invariant holds trivially as $0\le\psi(\mathbf{C}^D[h^0])\le 1$. Assume the invariant holds at the end of iteration $t-1\in\{0,\dots,T-1\}$; we shall prove that the invariant holds at the end of iteration $t$. In particular, we consider two cases at iteration $t$. In the first case, $\psi(\widehat{\mathbf{\Gamma}}^t)\ge\gamma^t$, leading to the assignments $\alpha^t=\gamma^t$, $\beta^t=\beta^{t-1}$, and $h^t=\widehat{g}^t$. We have from the definition of $\epsilon'$

$$\begin{aligned}\langle\mathbf{A}-\gamma^t\mathbf{B}, \mathbf{C}^D[\widehat{g}^t]\rangle) &\ge \langle\mathbf{A}-\gamma^t\mathbf{B},\widehat{\mathbf{\Gamma}}^t\rangle - \|\mathbf{A}-\gamma^t\mathbf{B}\|_1\epsilon'\\ &= \langle\mathbf{A},\widehat{\mathbf{\Gamma}}^t\rangle - \gamma^t\langle\mathbf{B},\widehat{\mathbf{\Gamma}}^t\rangle - \|\mathbf{A}-\gamma^t\mathbf{B}\|_1\epsilon'\\ &= \langle\mathbf{B},\widehat{\mathbf{\Gamma}}^t\rangle\big(\psi(\widehat{\mathbf{\Gamma}}^t)-\gamma^t\big) - \|\mathbf{A}-\gamma^t\mathbf{B}\|_1\epsilon'\\ &\ge 0 - \|\mathbf{A}-\gamma^t\mathbf{B}\|_1\epsilon'\\ &> -\|\mathbf{A}-\gamma^t\mathbf{B}\|_1(2\epsilon+\epsilon')\\ &\ge -(\|\mathbf{A}\|_1+\|\mathbf{B}\|_1)(2\epsilon+\epsilon'),\end{aligned}$$

where the third step follows from our case assumption that $\psi(\widehat{\mathbf{\Gamma}}^t)\ge\gamma^t$ and $\langle\mathbf{B},\widehat{\mathbf{\Gamma}}^t\rangle>0$, the fifth step follows from $\epsilon>0$, and the last step follows from triangle inequality and $\gamma^t\le\sup_{\mathbf{C}\in\mathcal{C}_D}\psi(\mathbf{C})\le 1$. The above inequality further gives us

$$\begin{aligned}\frac{\langle\mathbf{A},\mathbf{C}^D[\widehat{g}^t]\rangle}{\langle\mathbf{B},\mathbf{C}^D[\widehat{g}^t]\rangle} &> \gamma^t - \frac{\|\mathbf{A}\|_1+\|\mathbf{B}\|_1}{\langle\mathbf{B},\mathbf{C}^D[\widehat{g}^t]\rangle}(2\epsilon+\epsilon')\\ &\ge \gamma^t - \frac{1}{b}(\|\mathbf{A}\|_1+\|\mathbf{B}\|_1)(2\epsilon+\epsilon')\\ &= \gamma^t - \tau(2\epsilon+\epsilon')\\ &= \gamma^t - \tau\bar{\epsilon}\\ &= \alpha^t - \tau\bar{\epsilon},\end{aligned}$$

where the second step follows from $\min_{\mathbf{C}\in\mathcal{C}_D} \langle \mathbf{B}, \mathbf{C}\rangle > b$ and the last step follows from the assignment $\alpha^t = \gamma^t$. In other words,

$$\psi\big(\mathbf{C}^D[h^t]\big) \;=\; \psi(\mathbf{C}^D[\widehat{g}^t]) \;=\; \frac{\langle \mathbf{A},\, \mathbf{C}^D[\widehat{g}^t]\rangle}{\langle \mathbf{B},\, \mathbf{C}^D[\widehat{g}^t]\rangle} \;>\; \alpha^t \;-\; \tau\bar{\epsilon}.$$

Moreover, by our assumption that the invariant holds at the end of iteration $t-1$, we have

$$\beta^t \;+\; \tau\bar{\epsilon} \;=\; \beta^{t-1} \;+\; \tau\bar{\epsilon} \;\geq\; \max_{\mathbf{C}\in\mathcal{C}_D} \psi(\mathbf{C}) \;\geq\; \psi(\mathbf{C}^D[h^t]) \;>\; \alpha^t \;-\; \tau\bar{\epsilon}.$$

Thus under the first case, the invariant holds at the end of iteration $t$.

In the second case, $\psi(\widehat{\mathbf{\Gamma}}^t) < \gamma^t$ at iteration $t$, which would lead to the assignments $\alpha^t = \alpha^{t-1}$, $\beta^t = \gamma^t$, and $h^t = h^{t-1}$. Since the invariant is assumed to hold at the end of iteration $t-1$, we have

$$\alpha^t - \tau\bar{\epsilon} \;=\; \alpha^{t-1} - \tau\bar{\epsilon} \;\leq\; \psi(\mathbf{C}^D[h^{t-1}]) \;=\; \psi(\mathbf{C}^D[h^t]). \tag{13}$$

Next, recall that $\widehat{\mathbf{L}}^t \in [0,1]^{n\times n}$ is a scaled and translated version of $-(\mathbf{A} - \gamma^t\mathbf{B})$; clearly, there exists $c_t \in \mathbb{R}$ and $0 < a_t \leq 2\|\mathbf{A} - \gamma^t\mathbf{B}\|_\infty$ such that $\mathbf{A} - \gamma^t\mathbf{B} = c_t - a_t\mathbf{L}_t$. Then for $\mathbf{C}^* = \operatorname*{argmax}_{\mathbf{C}\in\mathcal{C}_D} \langle \mathbf{A} - \gamma^t\mathbf{B},\, \mathbf{C}\rangle$, we have

$$\begin{aligned}
\langle \mathbf{A} - \gamma^t\mathbf{B},\, \mathbf{C}^*\rangle &= c_t - a_t\langle \widehat{\mathbf{L}}^t,\, \mathbf{C}^*\rangle \\
&\leq c_t - a_t\langle \widehat{\mathbf{L}}^t, \mathbf{C}^D[h^t]\rangle \;+\; a_t\epsilon \\
&\leq c_t - a_t\langle \widehat{\mathbf{L}}^t, \mathbf{C}^D[h^t]\rangle \;+\; 2\|\mathbf{A} - \gamma^t\mathbf{B}\|_\infty\epsilon \\
&= \langle \mathbf{A} - \gamma^t\mathbf{B},\, \mathbf{C}^D[h^t]\rangle \;+\; 2\|\mathbf{A} - \gamma^t\mathbf{B}\|_\infty\epsilon \\
&\leq \langle \mathbf{A} - \gamma^t\mathbf{B},\, \widehat{\mathbf{\Gamma}}^t\rangle \;+\; \|\mathbf{A} - \gamma^t\mathbf{B}\|_1\epsilon' \;+\; 2\|\mathbf{A} - \gamma^t\mathbf{B}\|_\infty\epsilon \\
&= \langle \mathbf{A},\, \widehat{\mathbf{\Gamma}}^t\rangle - \gamma^t\langle \mathbf{B},\, \widehat{\mathbf{\Gamma}}^t\rangle \;+\; \|\mathbf{A} - \gamma^t\mathbf{B}\|_1\epsilon' \;+\; 2\|\mathbf{A} - \gamma^t\mathbf{B}\|_\infty\epsilon \\
&= \langle \mathbf{B},\, \widehat{\mathbf{\Gamma}}^t\rangle(\psi(\widehat{\mathbf{\Gamma}}^t) - \gamma^t) \;+\; \|\mathbf{A} - \gamma^t\mathbf{B}\|_1\epsilon' \;+\; 2\|\mathbf{A} - \gamma^t\mathbf{B}\|_\infty\epsilon \\
&\leq \langle \mathbf{B},\, \widehat{\mathbf{\Gamma}}^t\rangle(0) \;+\; \|\mathbf{A} - \gamma^t\mathbf{B}\|_1\epsilon' \;+\; 2\|\mathbf{A} - \gamma^t\mathbf{B}\|_\infty\epsilon \\
&\leq \|\mathbf{A} - \gamma^t\mathbf{B}\|_1(2\epsilon + \epsilon') \\
&\leq (\|\mathbf{A}\|_1 + \|\mathbf{B}\|_1)(2\epsilon + \epsilon'),
\end{aligned}$$

where the second step follows from Eq. (12), the third step uses $a_t \leq \|\mathbf{A} - \gamma^t\mathbf{B}\|_\infty$, the fifth step follows from the definition of $\epsilon'$ and Holder's inequality, the seventh step follows from our case assumption that $\psi(\widehat{\mathbf{\Gamma}}^t) \leq \gamma^t$ and $\langle \mathbf{B},\, \widehat{\mathbf{\Gamma}}^t\rangle > 0$, and the last step follows from triangle inequality and $\gamma^t \leq \sup_{\mathbf{C}\in\mathcal{C}_D} \psi(\mathbf{C}) \leq 1$. In particular, we have for all $\mathbf{C} \in \mathcal{C}_D$,

$$\langle \mathbf{A} - \gamma^t\mathbf{B},\, \mathbf{C}\rangle \;\leq\; (\|\mathbf{A}\|_1 + \|\mathbf{B}\|_1)(2\epsilon + \epsilon'),$$

or

$$\frac{\langle \mathbf{A},\, \mathbf{C}\rangle}{\langle \mathbf{B},\, \mathbf{C}\rangle} \;\leq\; \gamma^t \;+\; \frac{\|\mathbf{A}\|_1 + \|\mathbf{B}\|_1}{\langle \mathbf{B},\, \mathbf{C}\rangle}(2\epsilon + \epsilon').$$

Since $\min_{\mathbf{C}\in\mathcal{C}_D} \langle \mathbf{B},\, \mathbf{C}\rangle > b$, we have from the above, for all $\mathbf{C} \in \mathcal{C}_D$,

$$\psi(\mathbf{C}) \;\leq\; \gamma^t \;+\; \frac{1}{b}(\|\mathbf{A}\|_1 + \|\mathbf{B}\|_1)(2\epsilon + \epsilon') \;=\; \gamma^t + \tau\bar{\epsilon} \;=\; \beta^t + \tau\bar{\epsilon}.$$

In other words,

$$\sup_{\mathbf{C}\in\mathcal{C}_D} \psi(\mathbf{C}) \;\leq\; \beta^t + \tau\bar{\epsilon}.$$

By combining the above with Eq. (13), we can see that the invariant holds in iteration $t$ under this case as well. This completes the proof of the lemma. □

**Lemma 28 (Multiplicative Progress in Each Iteration of Algorithm 2).** *Let $\psi$ be as defined in Theorem 17. Then the following is true in each iteration $1 \leq t \leq T$ of Algorithm 2:*

$$\beta^t - \alpha^t \;=\; \frac{1}{2}\big(\beta^{t-1} - \alpha^{t-1}\big).$$

*Proof.* We consider two cases in each iteration of Algorithm 2. If in an iteration $t \in \{1, \ldots, T\}$, $\psi(\widehat{\mathbf{\Gamma}}^t) \geq \gamma^t$, leading to the assignment $\alpha^t = \gamma^t$, then

$$\begin{aligned}\beta^t - \alpha^t &= \beta^{t-1} - \gamma^t \\ &= \beta^{t-1} - \frac{\alpha^{t-1} + \beta^{t-1}}{2} \\ &= \frac{1}{2}\big(\beta^{t-1} - \alpha^{t-1}\big).\end{aligned}$$

On the other hand, if $\psi(\widehat{\mathbf{\Gamma}}^t) < \gamma^t$, leading to the assignment $\beta^t = \gamma^t$, then

$$\begin{aligned}\beta^t - \alpha^t &= \gamma^t - \alpha^{t-1} \\ &= \frac{\alpha^{t-1} + \beta^{t-1}}{2} - \alpha^{t-1} \\ &= \frac{1}{2}(\beta^{t-1} - \alpha^{t-1}).\end{aligned}$$

Thus in both cases, the statement of the lemma is seen to hold. □

We now prove Theorem 17.

*Proof of Theorem 17.* For the classifier $h_S^{\text{BS}} = h^T$ output by Algorithm 2 after $T$ iterations, we have from Lemma 27

$$\begin{aligned}\mathcal{P}_D^* \,-\, \mathcal{P}_D[h_S^{\text{BS}}] &= \sup_{\mathbf{C} \in \mathcal{C}_D} \psi(\mathbf{C}) \,-\, \psi\big(\mathbf{C}^D[h^T]\big) \\ &< (\beta^t \,+\, \tau\bar{\epsilon}) \,-\, (\alpha^t \,-\, \tau\bar{\epsilon}) \\ &= \beta^t - \alpha^t \,+\, 2\tau\bar{\epsilon} \\ &\leq 2^{-T}\big(\beta^0 - \alpha^0\big) \,+\, 2\tau\bar{\epsilon} \\ &= 2^{-T}\big(1 - 0\big) \,+\, 2\tau\bar{\epsilon} \\ &= 2^{-T} \,+\, 2\tau\bar{\epsilon},\end{aligned}$$

where $\bar{\epsilon}$ is as defined in Lemma 27; the fifth step above follows from Lemma 28. Setting $T = \kappa m$ thus gives us

$$\mathcal{P}_D^* \,-\, \mathcal{P}_D[h_S^{\text{BS}}] \;\leq\; 2\tau \mathbf{E}_X\big[\big\|\widehat{\boldsymbol{\eta}}(X) - \boldsymbol{\eta}(X)\big\|_1\big] \,+\, 2\tau \sup_{h \in \mathcal{H}_{\boldsymbol{\eta}}} \big\|\mathbf{C}^D[h] \,-\, \widehat{\mathbf{C}}^{S_2}[h]\big\|_\infty \,+\, 2^{-\kappa m}.$$

By an application Lemma 15 to the second term in the right-hand side of the above inequality (noting that $|S_2| = \lfloor m/2 \rfloor$), we then have for any $\delta > 0$, with probability at least $1 - \delta$,

$$\mathcal{P}_D^* \,-\, \mathcal{P}_D[\widehat{h}_S^{\text{BS}}] \;\leq\; 2\tau \mathbf{E}_X\big[\big\|\widehat{\boldsymbol{\eta}}(X) - \boldsymbol{\eta}(X)\big\|_1\big] \,+\, 2\sqrt{2}C\tau\sqrt{\frac{n^2 \log(n) \log(m) + \log(n^2/\delta)}{m}} \,+\, 2^{-\kappa m},$$

for a distribution-independent constant $C > 0$. □

### C.5. Extending Algorithm 1 to Non-Smooth Performance Measures

In Section 5, we showed that Algorithm 1 was consistent for any concave smooth performance measure (see Theorem 16). We now extend this result to concave performance measures for which the associated $\psi$ is non-smooth (but differentiable); these include the G-mean, H-mean and Q-mean performance measures in 1. In particular, for these performance measures, we prescribe that Algorithm 1 be applied to a suitable smooth approximation to $\psi$; if the quality of this approximation improves with the size of the given training sample (at an appropriate rate), then the resulting algorithm can be shown to be consistent for the original performance measure.

**Theorem 29.** *Let $\psi : [0,1]^{n\times n}\to\mathbb{R}_+$ be such that for any $\rho \in \mathbb{R}_+$, there exists $\psi_\rho : [0,1]^{n\times n}\to\mathbb{R}_+$ which is concave over $\mathcal{C}_D$, $L_\rho$ Lipschitz and $\beta_\rho$ smooth w.r.t. the $\ell_1$ norm with*

$$\sup_{\mathbf{C}\in\mathcal{C}_D} |\psi(\mathbf{C}) - \psi_\rho(\mathbf{C})| \leq \theta(\rho),$$

*for some strictly increasing function $\theta : \mathbb{R}_+\to\mathbb{R}_+$. Let $S = (S_1, S_2) \in (\mathcal{X}\times[n])^m$ be a training sample drawn i.i.d. from $D$. Further, let $\widehat{\boldsymbol{\eta}} : \mathcal{X}\to\Delta_n$ be the* CPE *model learned from $S_1$ in Algorithm 1 and $h_S^{\text{FW},\rho} : \mathcal{X}\to\Delta_n$ be the classifier obtained after $\kappa m$ iterations by Algorithm 1 when run for the performance measure $\psi_\rho$. Then for any $\delta \in (0,1]$, with probability at least $1-\delta$ (over draw of $S$ from $D^m$),*

$$\mathcal{P}_D^{\psi,*} - \mathcal{P}_D^{\psi}[h_S^{\text{FW},\rho}] \leq 4L_\rho \mathbf{E}_X\left[\left\|\widehat{\boldsymbol{\eta}}(X) - \boldsymbol{\eta}(X)\right\|_1\right] + 4\sqrt{2}\beta_\rho n^2 C\sqrt{\frac{n^2\log(n)\log(m)+\log(n^2/\delta)}{m}} + \frac{8\beta_\rho}{\kappa m+2} + 2\theta(\rho),$$

*where $C > 0$ is a distribution-independent constant.*

*Proof.* From Theorem 16 we have that

$$\mathcal{P}_D^{\psi_\rho,*} - \mathcal{P}_D^{\psi_\rho}[h_S^{\text{FW},\rho}] \leq 4L_\rho \mathbf{E}_X\left[\left\|\widehat{\boldsymbol{\eta}}(X) - \boldsymbol{\eta}(X)\right\|_1\right] + 4\sqrt{2}\beta_\rho n^2 C\sqrt{\frac{n^2\log(n)\log(m)+\log(n^2/\delta)}{m}} + \frac{8\beta_\rho}{\kappa m+2}. \quad (14)$$

For simplicity assume that the $\psi$-optimal classifier exists; the proof can be easily extended when this is not the case. Let $h^* : \mathcal{X}\to\Delta_n$ be a $\psi$-optimal classifier; note that this classifier need not be $\psi_\rho$-optimal. We then have that

$$\begin{aligned}
&\mathcal{P}_D^{\psi,*} - \mathcal{P}_D^{\psi}[h_S^{\text{FW},\rho}] \\
&= \psi(\mathbf{C}^D[h^*]) - \psi(\mathbf{C}^D[h_S^{\text{FW},\rho}]) \\
&\leq \psi_\rho(\mathbf{C}^D[h^*]) - \psi_\rho(\mathbf{C}^D[h_S^{\text{FW},\rho}]) + 2\theta(\rho) \\
&= \mathcal{P}_D^{\psi_\rho}[h^*] - \mathcal{P}_D^{\psi_\rho}[h_S^{\text{FW},\rho}] + 2\theta(\rho) \\
&\leq \mathcal{P}_D^{\psi_\rho,*} - \mathcal{P}_D^{\psi_\rho}[h_S^{\text{FW},\rho}] + 2\theta(\rho) \\
&\leq 4L_\rho \mathbf{E}_X\left[\left\|\widehat{\boldsymbol{\eta}}(X) - \boldsymbol{\eta}(X)\right\|_1\right] + 4\sqrt{2}\beta_\rho n^2 C\sqrt{\frac{n^2\log(n)\log(m)+\log(n^2/\delta)}{m}} + \frac{8\beta_\rho}{\kappa m+2} + 2\theta(\rho),
\end{aligned}$$

where the second step follows from our assumption that $\sup_{\mathbf{C}\in\mathcal{C}_D} |\psi(\mathbf{C}) - \psi_\rho(\mathbf{C})| \leq \theta(\rho)$, and the fifth step follows from the definition of $\mathcal{P}^{\psi_\rho,*}$ and the last step uses Eq. (14). This completes the proof. □

We note that for each of G-mean, H-mean and Q-mean, one can construct a Lipschitz smooth approximation $\psi_\rho$ as required in the above theorem. Now, suppose the CPE algorithm in Algorithm 1 is such that the class probability estimation error term in the theorem $\mathbf{E}_X\left[\left\|\widehat{\boldsymbol{\eta}}(X) - \boldsymbol{\eta}(X)\right\|_1\right] \xrightarrow{P} 0$ (as the number of training examples $m\to\infty$). Then for each of the given performance measures, one can allow the parameter $\rho$ (that determines the approximation quality of $\psi_\rho$) to go to 0 as $m\to\infty$ (at appropriate rate), so that the right-hand side of the bound in the theorem goes to 0 (as $m\to\infty$), implying that Algorithm 1 is $\psi$-consistent. We postpone the details to a longer version of the paper.

## D. Supplementary Material for Section 6 (Experiments)

### D.1. Computation of Class Probability Function for Distribution Considered in Synthetic Data Experiments

We provide the calculations for the class probability function for the distribution considered in synthetic data experiments in Section 6. We present this for a more general distribution over $\mathbb{R}^d\times[n]$, where for each class $i$, the class prior probability is $\pi_i$ and the class conditional distribution is a Gaussian distribution with mean $\mu_i \in \mathbb{R}^d$ and the same (symmetric positive semidefinite) covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$. We shall denote the pdf for the Gaussian corresponding to class $i$ as $f_i(x) = \frac{1}{\sqrt{(2\pi)^d|\Sigma|}}\exp\left(-\frac{1}{2}(x-\mu_i)^\top\Sigma^{-1}(x-\mu_i)\right)$. The class probability function for this distribution is then given by

$$\begin{aligned}
\eta_i(x) &= \mathbf{P}(Y = i \,|\, X = x) \\
&= \frac{\pi_i f_i(x)}{\sum_{j=1}^n \pi_j f_j(x)}
\end{aligned}$$

*Table 6.* Data sets used in experiments in Sections 6.2–6.4.

| | Data set | # instances | # features | # classes |
|---|---|---|---|---|
| **UCI** | car | 1728 | 21 | 4 |
| | pageblocks | 5473 | 10 | 5 |
| | glass | 214 | 9 | 6 |
| | satimage | 6435 | 36 | 6 |
| | covtype | 581012 | 54 | 7 |
| | yeast | 1484 | 8 | 10 |
| | abalone | 4177 | 10 | 12 |
| **IR** | cora | 2708 | 1433 | 4 |
| | news20 | 12199 | 61188 | 4 |
| | rcv1 | 15564 | 47236 | 11 |

$$
\begin{aligned}
&= \frac{\exp\left(-\frac{1}{2}(x-\mu_i)^\top\Sigma^{-1}(x-\mu_i) + \ln\pi_i\right)}{\sum_{j=1}^n \exp\left(-\frac{1}{2}(x-\mu_j)^\top\Sigma^{-1}(x-\mu_j) + \ln\pi_j\right)} \\
&= \frac{\exp\left(\mu_i^\top\Sigma^{-1}x - \frac{1}{2}\mu_i^\top\Sigma^{-1}\mu_i + \ln\pi_i\right)\exp\left(-\frac{1}{2}x^\top\Sigma^{-1}x\right)}{\sum_{j=1}^n \exp\left(\mu_j^\top\Sigma^{-1}x - \frac{1}{2}\mu_j^\top\Sigma^{-1}\mu_j + \ln\pi_j\right)\exp\left(-\frac{1}{2}x^\top\Sigma^{-1}x\right)} \\
&= \frac{\exp(w_i^\top x + b_i)}{\sum_{j=1}^n \exp(w_j^\top x + b_j)},
\end{aligned}
$$

where $w_i = \Sigma^{-1}\mu_i$ and $b_i = -\frac{1}{2}\mu_i^\top\Sigma^{-1}\mu_i + \ln\pi_i$. Clearly, the class probability function for the distribution considered can be obtained as a softmax of a linear function.

### D.2. Additional Experimental Details/Results

In all our experiments, the regularization parameter for each algorithm was chosen from the range $\{10^{-4}, \ldots, 10^4\}$ using a held-out portion of the training set.

**Synthetic data experiments.** Since the distribution used to generate synthetic data and the four performance measures considered satisfy the condition in Theorem 13, the optimal classifier for each performance measure can be obtained by computing the $\psi^{\mathbf{L}^*}$-optimal classifier for some loss matrix $\mathbf{L}^* \in [0,1]^{n\times n}$; we have a similar characterization for micro $F_1$ using Theorem 13. In our experiments, we computed the optimal performance for a given performance measure by performing a search over a large range of $n \times n$ loss matrices $\mathbf{L}$, used the true conditional class probability to compute a $\psi^{\mathbf{L}}$-optimal classifier for each such $\mathbf{L}$ (see Proposition 6), and chose among these classifiers the one which gave the highest performance value (on a large sample drawn from the distribution). Moreover, since the class probability function here is a softmax of linear functions, it follows that the Bayes optimal performance is also achieved by a linear classification model, and therefore learning a linear model suffices to achieve consistency; we therefore learn a linear classification model in all experiments. Also, recall that Algorithm 1 outputs a randomized classifier, while Algorithm 2 outputs a deterministic classifier. In our experimental results, the $\psi$-performance of a randomized classifier was evaluated using the 'expected' (empirical) confusion matrix of the deterministic classifiers in its support.

**Real data experiments.** All real data sets used in our experiments have been listed in Table 6. The version of the CoRA data set used in our experiments was obtained from `http://membres-lig.imag.fr/grimal/data.html`. The 20 Newsgroup data was obtained from `http://qwone.com/~jason/20Newsgroups/`. For the RCV1 data, we used a preprocessed version obtained from `http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/`. For each of the UCI data sets used in our experiments, the training set was normalized to 0 mean and unit variance, and this transformation was applied on the test set. Table 7–9 contains results on UCI data sets not provided in Section 6. Table 10–12 contains training times for Algorithm 1 (applied to the G-mean, H-mean and Q-mean measures) and the baseline SVM$^{\text{perf}}$ and 0-1 logistic regression methods on all UCI data sets; in each case, the symbol $\times$ against SVM$^{\text{perf}}$ indicates the method did not complete after 96 hrs.

**Implementation details.** The proposed Frank-Wolfe based and bisection based algorithms were implemented in MATLAB; in order to learn a CPE model in these algorithms, we used the multiclass logistic regression solver provided in `http://www.cs.ubc.ca/~schmidtm/Software/minFunc.html` for the experiments on the synthetic and UCI data sets, and the liblinear logistic regression implementation provided in `www.csie.ntu.edu.tw/~cjlin/`

| | car | pageblocks | glass | satimage | covtype | yeast | abalone |
|---|---|---|---|---|---|---|---|
| Frank-Wolfe (GM) | **0.945** | **0.908** | **0.680** | **0.843** | **0.695** | **0.448** | **0.223** |
| SVM$^{\text{perf}}$ (GM) | 0.792 | 0.796 | 0.431 | × | × | × | × |
| LogReg (0-1) | 0.911 | 0.691 | 0.146 | 0.779 | 0.692 | 0.000 | 0.000 |

*Table 7.* Results for G-mean on UCI data sets.

| | car | pageblocks | glass | satimage | covtype | yeast | abalone |
|---|---|---|---|---|---|---|---|
| Frank-Wolfe (HM) | **0.945** | **0.904** | **0.632** | **0.836** | **0.686** | **0.412** | **0.197** |
| SVM$^{\text{perf}}$ (HM) | 0.880 | 0.574 | 0.381 | × | × | × | × |
| LogReg (0-1) | 0.909 | 0.631 | 0.143 | 0.731 | 0.679 | 0.000 | 0.000 |

*Table 8.* Results for H-mean on UCI data sets.

| | car | pageblocks | glass | satimage | covtype | yeast | abalone |
|---|---|---|---|---|---|---|---|
| Frank-Wolfe (QM) | **0.930** | **0.877** | **0.613** | **0.821** | **0.685** | **0.510** | **0.247** |
| SVM$^{\text{perf}}$ (QM) | 0.909 | 0.651 | 0.481 | × | × | × | × |
| LogReg (0-1) | 0.898 | 0.660 | 0.490 | 0.725 | 0.675 | 0.473 | 0.223 |

*Table 9.* Results for Q-mean on UCI data sets.

| | car | pageblocks | glass | satimage | covtype | yeast | abalone |
|---|---|---|---|---|---|---|---|
| Frank-Wolfe (GM) | 1.96 | 5.89 | 0.27 | 9.66 | 139.60 | 1.68 | 7.31 |
| SVM$^{\text{perf}}$ (GM) | 8327.54 | 63667.67 | 1302.84 | × | × | × | × |
| LogReg (0-1) | 0.59 | 1.70 | 0.07 | 4.48 | 106.27 | 0.40 | 3.84 |

*Table 10.* Training time (in secs) for G-mean on UCI data sets.

| | car | pageblocks | glass | satimage | covtype | yeast | abalone |
|---|---|---|---|---|---|---|---|
| Frank-Wolfe (HM) | 1.96 | 5.85 | 0.26 | 9.02 | 125.30 | 1.69 | 7.14 |
| SVM$^{\text{perf}}$ (HM) | 3342.08 | 35836.87 | 108.80 | × | × | × | × |
| LogReg (0-1) | 0.57 | 1.55 | 0.07 | 4.78 | 127.12 | 0.38 | 4.07 |

*Table 11.* Training time (in secs) for H-mean on UCI data sets.

| | car | pageblocks | glass | satimage | covtype | yeast | abalone |
|---|---|---|---|---|---|---|---|
| Frank-Wolfe (QM) | 1.93 | 6.11 | 0.27 | 9.00 | 134.85 | 1.65 | 7.29 |
| SVM$^{\text{perf}}$ (QM) | 6795.87 | 54803.42 | 158.48 | × | × | × | × |
| LogReg (0-1) | 0.61 | 1.79 | 0.07 | 4.72 | 120.60 | 0.43 | 3.84 |

*Table 12.* Training time (in secs) for Q-mean on UCI data sets.

`liblinear` for the experiments on IR data. All run-time experiments were run on Intel Xeon quad-core machines (2.66 GHz, 12 MB cache) with 16 GB RAM.

We implemented SVM$^{\text{perf}}$ using a publicly available structural SVM API[9]. The SVM$^{\text{perf}}$ method (proposed originally for binary performance measures (Joachims, 2005)) uses a cutting plane solver where computing the most-violated constraint requires a search over of all valid confusion matrices for the given training sample. In the case of the G-mean, H-mean and Q-mean measures, this search can be restricted to the diagonal entries of the confusion matrix, but will still require (in the worst case) time exponential in the number of classes; in the case of the micro $F_1$, this search is more expensive and will involve searching over $3n-3$ entries of the confusion matrix. While we use an exact implementation of SVM$^{\text{perf}}$ for these three performance measures, for the micro $F_1$-measure, we use a version that optimizes an approximation to the micro $F_1$ (in particular, optimizes the variant of micro $F_1$ analyzed by (Parambath et al., 2014)) and requires fewer computations. The tolerance parameter for the cutting-plane method in SVM$^{\text{perf}}$ was set to 0.01 for all experiments except on the Pageblocks and CoRA data sets, where the tolerance was set to 0.1 to enable faster run-time.

[9] http://www.cs.cornell.edu/people/tj/svm_light/svm_struct.html