A. Proofs from Section 2

A.1. Proof of Proposition 1

For the sake of readability, throughout the proof we abbreviate $\Phi = \Phi(u,v,p)$, $\Phi' = \Phi(u',v',p')$, and denote $\Delta u = u - u'$, $\Delta v = v - v'$, $\Delta p = p - p'$. In this notation, proving p-Lipschitzness for metric Φ amounts to showing that:

$$|\Phi - \Phi'| \le U_p|\Delta u| + V_p|\Delta v| + P_p|\Delta p|,$$

for constants U_p, V_p, P_p, which may only depend on p.

The following fact is going to be very useful in proving p-Lipschitzness. If the metric is of the rational form: $\Phi(u,v,p) = \frac{A(u,v,p)}{B(u,v,p)} + C$, where C is some constant, $B(u,v,p) \ge G_p$ for some positive constant G_p (which may depend on p), and $|\Phi(u,v,p)| \le \Phi_{\max}$ for some constant $\Phi_{\max}$, it suffices to check p-Lipschitzness of numerator and denominator separately. Indeed, using shorthand notation $A = A(u,v,p)$, $A' = A(u',v',p')$, and similarly for B, B':

$$\begin{aligned}\Phi - \Phi' &= \frac{A - \frac{A'}{B'}B}{B} = \frac{A - A' + \frac{A'}{B'}B' - \frac{A'}{B'}B}{B} \\ &= \frac{A - A'}{B} + \frac{A'}{B'}\frac{B - B'}{B},\end{aligned}$$

hence:

$$|\Phi - \Phi'| \le \frac{|A - A'|}{G_p} + \frac{\Phi_{\max}}{G_p}|B' - B|.$$

a) *Accuracy* $\Phi(u,v,p) = 1 - v - p + 2u$. We have:

$$\Phi - \Phi' \le 2\Delta u - \Delta v - \Delta p,$$

so that by triangle inequality:

$$|\Phi - \Phi'| \le 2|\Delta u| + |\Delta v| + |\Delta p|.$$

Hence, the statement follows with $U_p = 2$, $V_p = P_p = 1$.

b) *AM* $\Phi(u,v,p) = 1 - \frac{vp-u}{2p(1-p)}$. We can use the result on the rational metric by noting that $A(u,v,p) = u - vp$, $B(u,v,p) = B(p) = 2p(1-p)$, $C = 1$, $\Phi_{\max} = 1$, $G_p = 2p(1-p)$. We can now check the p-Lipschitzness of A and B separately:

$$\begin{aligned}A - A' &= u - vp - u' + v'p' \\ &= \Delta u + (vp' - vp) + (v'p' - vp') \\ &= \Delta u - v\Delta p - p'\Delta v,\end{aligned}$$

and since $|v| \le 1$, $|p'| \le 1$, p-Lipschitzness follows from triangle inequality. For the denominator,

$$\begin{aligned}B - B' &= 2p(1-p) - 2p'(1-p') \\ &= 2(p - p') + 2(p'^2 - p^2) \\ &= 2(1 - p' - p)(p - p'),\end{aligned}$$

so that $|B - B'| \le 2|\Delta p|$.

c) *Jaccard similarity* $\Phi(u,v,p) = \frac{u}{p+v-u}$. Follows from the rational form of the metric, since $A(u,v,p) = u$, $B(u,v,p) = p + v - u$, $C = 0$, $\Phi_{\max} = 1$, $G_p = p$, and the p-Lipschitzness of $A(u,v,p)$ and $B(u,v,p)$ is trivial to show by the triangle inequality.

d) *G-mean* $\Phi(u,v,p) = \frac{u(1-v-p+u)}{p(1-p)}$. Exploiting the rational form of the metric, we have $A(u,v,p) = u(1 - v - p + u)$, $B(u,v,p) = p(1-p)$, $C = 0$, $\Phi_{\max} = 1$, $G_p = p(1-p)$. The p-Lipschitzness of B was shown above for AM measure. As for A:

$$\begin{aligned}A - A' &= (1 - v - p + u)(u - u') \\ &\quad + u'(u - p - v - u' - p' - v') \\ &= (1 - v - p + u)\Delta u + u'(\Delta u - \Delta v - \Delta p),\end{aligned}$$

and hence the p-Lipschitzness follows by triangle inequality and the fact that $|1 - v - p + u| \le 2$ and $|u'| \le 1$.

e) *AUC* $\frac{(v-u)(p-u)}{p(1-p)}$. Exploiting the rational form of the metric, we have $A(u,v,p) = (v-u)(p-u)$ and $B(u,v,p) = p(1-p)$. The p-Lipschitzness of B was shown above for AM measure; as for A:

$$\begin{aligned}A - A' &= (v-u)(p-u) - (v'-u')(p-u) \\ &\quad + (v'-u')(p-u) - (v'-u')(p'-u') \\ &= (\Delta v - \Delta u)(p-u) + (v'-u')(\Delta p - \Delta u),\end{aligned}$$

and hence the p-Lipschitzness follows by triangle inequality and the fact that $|p-u| \le 1$ and $|v'-u'| \le 1$.

f) *Linear-fractional metric* of the form:

$$\Phi(u,v,p) = \frac{a_1 + a_2 u + a_3 v + a_4 p}{b_1 + b_2 u + b_3 v + b_4 p},$$

as long as the denominator is bounded from below by some positive constant G_p. This follows immediately from the rational form of the metric, as the numerator $A(u,v,p)$ and denominator $B(u,v,p)$ are linear functions of (u,v,p), so showing p-Lipschitzness of $A(u,v,p)$ and $B(u,v,p)$ is straightforward.

B. Proofs from Section 3.1

B.1. Proof of Lemma 1

We fix classifier h and use a shorthand notation $u, v, \widehat{u}, \widehat{v}$ to denote $u(h), v(h), \widehat{u}(h), \widehat{v}(h)$. Due to the Lipschitz assumption:

$$|\Phi(u,v,p) - \Phi(\widehat{u},\widehat{v},\widehat{p})| \le U_p|u - \widehat{u}| + V_p|v - \widehat{v}| + P_p|p - \widehat{p}|.$$

Fixing $\boldsymbol{x} = (x_1, \ldots, x_n)$ and taking expectation with respect to $\boldsymbol{y} = (y_1, \ldots, y_n)$ conditioned on $\boldsymbol{x}$, we have:

$$\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[|\Phi(u,v,p) - \Phi(\widehat{u},\widehat{v},\widehat{p})|\right] \\ \leq U_p \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[|u-\widehat{u}|\right] + V_p|v-\widehat{v}| + P_p \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[|p-\widehat{p}|\right].$$

Denote:

$$\widetilde{p} = \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[\widehat{p}\right] = \frac{1}{n}\sum_{i=1}^n \eta(x_i),$$

$$\widetilde{u} = \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[\widehat{u}\right] = \frac{1}{n}\sum_{i=1}^n h(x_i)\eta(x_i)$$

We have:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[|p-\widehat{p}|\right] &= \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[|p-\widetilde{p}+\widetilde{p}-\widehat{p}|\right] \\ &\leq |p-\widetilde{p}| + \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[|\widetilde{p}-\widehat{p}|\right] \\ &= |p-\widetilde{p}| + \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[\sqrt{(\widetilde{p}-\widehat{p})^2}\right] \\ &\leq |p-\widetilde{p}| + \sqrt{\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[(\widetilde{p}-\widehat{p})^2\right]} \\ &= |p-\widetilde{p}| + \sqrt{\mathrm{Var}_{\boldsymbol{y}|\boldsymbol{x}}(\widehat{p})} \leq |p-\widetilde{p}| + \sqrt{\frac{1}{4n}},\end{aligned}$$

where the second inequality follows from Jensen's inequality applied to a concave function $x \mapsto \sqrt{x}$. In an analogous way, one can show that:

$$\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[|u-\widehat{u}|\right] \leq |u-\widetilde{u}| + \sqrt{\frac{u}{4n}} \leq |u-\widetilde{u}| + \sqrt{\frac{1}{4n}}.$$

Furthermore, using the convexity of the absolute value function, Jensen's inequality implies:

$$\left|\Phi(u,v,p) - \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[\Phi(\widehat{u},\widehat{v},\widehat{p})\right]\right| \\ \leq \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[|\Phi(u,v,p) - \Phi(\widehat{u},\widehat{v},\widehat{p})|\right],$$

so that:

$$\left|\Phi(u,v,p) - \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[\Phi(\widehat{u},\widehat{v},\widehat{p})\right]\right| \leq U_p|u-\widetilde{u}| + V_p|v-\widehat{v}| \\ + P_p|p-\widetilde{p}| + \frac{U_p+V_p}{2\sqrt{n}}.$$

We will now show that under the class of thresholded functions $\mathcal{H}$ specified in the statement of the theorem to which h belongs, all the terms on the right-hand side are well controlled. The rest of the proof follows in a straightforward way from Hoeffding's inequality and Vapnik-Chervonenkis bounds, except for minor, technical details, which are included for completeness.

We first apply Hoeffding's inequality to say that with probability at least $1-\delta/2$,

$$|p-\widetilde{p}| \leq \sqrt{\frac{\log\frac{4}{\delta}}{2n}}.$$

Similarly, using standard Rademacher complexity arguments (see, e.g. Mohri et al., 2012), we have, uniformly over all $h \in \mathcal{H}$, with probability $1-\delta/4$,

$$|v-\widehat{v}| \leq 2\mathbb{E}_{\boldsymbol{x}}\left[\mathcal{R}_n(\mathcal{H})\right] + \sqrt{\frac{\log\frac{4}{\delta}}{2n}},$$

and similarly, with probability $1-\delta/4$,

$$|u-\widetilde{u}| \leq 2\mathbb{E}_{\boldsymbol{x}}\left[\mathcal{R}_n(\mathcal{H}_\eta)\right] + \sqrt{\frac{\log\frac{4}{\delta}}{2n}},$$

where $\mathcal{H}_\eta = \{h \cdot \eta \colon h \in \mathcal{H}\}$, and:

$$\mathcal{R}_n(\mathcal{H}) = \mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\Big|\sum_{i=1}^n \sigma_i h(x_i)\Big|\right]$$

is the Rademacher complexity[6] of $\mathcal{H}$. Furthermore, if we let $z_i \in \{-1,1\}$, $i=1,\ldots,n$, with $\Pr(z_i = 1) = \frac{1+\eta(x_i)}{2}$, so that $\mathbb{E}\left[z_i\right] = \eta(x_i)$, we have:

$$\sum_{i=1}^n \sigma_i h(x_i)\eta(x_i) = \mathbb{E}_{\boldsymbol{z}}\Big[\sum_{i=1}^n \sigma_i h(x_i) z_i\Big],$$

so that:

$$\begin{aligned}\mathcal{R}_n(\mathcal{H}_\eta) &= \mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}}\frac{1}{n}\Big|\mathbb{E}_{\boldsymbol{z}}\Big[\sum_{i=1}^n \sigma_i h(x_i) z_i\Big]\Big|\right] \\ &\leq \mathbb{E}_{\sigma,\boldsymbol{z}}\left[\sup_{h\in\mathcal{H}}\frac{1}{n}\Big|\sum_{i=1}^n \sigma_i h(x_i) z_i\Big|\right] \\ &= \mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}}\frac{1}{n}\Big|\sum_{i=1}^n \sigma_i h(x_i)\Big|\right] = \mathcal{R}_n(\mathcal{H}),\end{aligned}$$

where the inequality is due to Jensen's inequality applied to convex functions $|\cdot|$ and $\sup\{\cdot\}$, and the second equality is due to the fact that $\sigma_i z_i$ and σ_i are distributed in the same way.

Thus choosing $L_p = \max\{U_p, V_p, P_p\}$, with probability $1-\delta$, uniformly over all $h \in \mathcal{H}$,

$$\left|\Phi(u,v,p) - \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[\Phi(\widehat{u},\widehat{v},\widehat{p})\right]\right| \leq 4L_p\mathbb{E}_{\boldsymbol{x}}\left[\mathcal{R}_n(\mathcal{H})\right] \\ + 3L_p\sqrt{\frac{\log\frac{4}{\delta}}{2n}} + \frac{L_p}{\sqrt{n}}.$$

Now, if $\mathcal{H}$ is the class of threshold functions on η, its growth function (Mohri et al., 2012) is equal to $m+1$, and thus we have[7]:

$$\mathcal{R}_n(\mathcal{H}) \leq \sqrt{\frac{2\log(n+1)}{n}},$$

[6] Variables σ_i, $i=1,\ldots,n$, are i.i.d. Rademacher variables distributed according to $\mathbb{P}(\sigma_i = 1) = \mathbb{P}(\sigma_i = -1) = \frac{1}{2}$.

[7] We could alternatively use the fact that VC-dimension of $\mathcal{H}$ is 1, which would give a bound with $\log(n+1)$ replaced by $1 + \log(n)$.

so that with probability $1-\delta$, uniformly over all $h \in \mathcal{H}$, we get the bound in the statement of the theorem. The proof is complete.

Lower bound. The dependence $\tilde{O}(1/\sqrt{n})$ on the sample size stated in Lemma 1 cannot be improved in general. To see this, take a metric $\Phi(u,v,p) = u$, p-Lipschitzness of which is trivial to show. Choose $h(x) = 1$ for all x. Then, $u(h) = p$, while $\widehat{u}(h) = \frac{1}{n}\sum_{i=1}^n y_i$. Hence, $\left|\Phi(u,v,p) - \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[\Phi(\widehat{u},\widehat{v},\widehat{p})\right]\right| = |p - \widetilde{p}|$, where $\widetilde{p} = \frac{1}{n}\sum_{i=1}^n \eta(x_i)$ and $\mathbb{E}_{\boldsymbol{x}}\left[\widetilde{p}\right] = p$. Assume that $\eta(x)$ follows a binomial distribution with $\mathbb{P}(\eta(x) = 1) = \mathbb{P}(\eta(x) = 0) = \frac{1}{2}$. Denote $|p - \widetilde{p}|$ by Z. By Khinchine inequality, $\mathbb{E}\left[Z\right] \geq 2c\sqrt{\mathbb{E}\left[Z^2\right]} = c/\sqrt{n}$ for some constant $c > 0$. Furthermore, by Paley-Zygmund inequality $\mathbb{P}(Z > \mathbb{E}\left[Z\right]/2) \geq \frac{(\mathbb{E}[Z])^2}{4\mathbb{E}[Z^2]} \geq c^2$. Hence, with constant probability,

$$\left|\Phi(u,v,p) - \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[\Phi(\widehat{u},\widehat{v},\widehat{p})\right]\right| \geq \frac{c}{2\sqrt{n}},$$

for some $c > 0$, which shows that the rate $\tilde{O}(1/\sqrt{n})$ cannot be improved.

B.2. Proof of Theorem 1

First, note that for a given $\mathbb{P}$, p-Lipschitzness implies that $\Phi(u,v,p)$ is continuous as a function of (u,v). Let $\mathcal{H} = \{h_\eta \mid h_\eta = \mathbb{1}_{\eta(x)\geq\eta}, \eta \in [0,1]\}$ be the set of binary threshold functions on $\eta(x)$. By Assumption 1, $u(h_\eta)$ and $v(h_\eta)$ are continuous in the threshold η, and hence the maximizer of $\Phi(u,v,p)$ over $\mathcal{H}$ exists due to compactness of the domain of η. The existence of the maximizer, together with Assumption 1 and TP monotonicity implies by (Narasimhan et al., 2014a, Lemma 11) that $h^*_{\mathrm{PU}} \in \mathcal{H}$, i.e. the optimal PU classifier is a threshold function.[8].

For any given $\boldsymbol{x} = (x_1,\ldots,x_n)$, let $h^*_{\mathrm{ETU}}(\boldsymbol{x})$ be the optimal ETU classifier. By TP monotonicity of Ψ, (Natarajan et al., 2016, Theorem 1) implies that $h^*_{\mathrm{ETU}}(\boldsymbol{x})$ satisfies:

$$\begin{aligned}\max_{i=1,\ldots,n} &\{\eta(x_i) \colon h^*_{\mathrm{ETU}}(x_i) = 0\} \\ &\leq \min_{i=1,\ldots,n}\{\eta(x_i) \colon h^*_{\mathrm{ETU}}(x_i) = 1\}.\end{aligned}$$

However, by Assumption 1, $\eta(x_i) \neq \eta(x_j)$ for all $i \neq j$ with probability one, so that the condition above is satisfied with strict inequality, and hence there exists τ^*, which is between $\max\{\eta(x_i) \colon h^*_{\mathrm{ETU}}(x_i) = 0\}$ and $\min\{\eta(x_i) \colon h^*_{\mathrm{ETU}}(x_i) = 1\}$. This means that $h^*_{\mathrm{ETU}}(\boldsymbol{x})$ is a threshold function on $\eta(x)$ with threshold τ^*, i.e. $h^*_{\mathrm{ETU}} \in \mathcal{H}$.

[8] Lemma 11 of Narasimhan et al. (2014a) requires that the PU maximizer within $\mathcal{H}$ is h_η for some $\eta \in (0,1)$. However, we do not impose this constraint here because the lemma can easily be extended to the case $\eta \in [0,1]$ under our assumption that $\eta(x)$ has a density over $[0,1]$.

To conclude, with probability one, $h^*_{\mathrm{ETU}}(\boldsymbol{x}), h^*_{\mathrm{PU}} \in \mathcal{H}$.

Now, define $\epsilon/2 = 4L_p\sqrt{\frac{2\log(n+1)}{n}} + 3L_p\sqrt{\frac{\log\frac{4}{\delta}}{2n}} + \frac{L_p}{\sqrt{n}}$. Then, with probability $1-\delta$ (over the random choice of $\boldsymbol{x}$),

$$\begin{aligned}\Phi(u(&h^*_{\mathrm{ETU}}(\boldsymbol{x})), v(h^*_{\mathrm{ETU}}(\boldsymbol{x})), p) \\ &\leq \Phi(u(h^*_{\mathrm{PU}}), v(h^*_{\mathrm{PU}}), p) \\ &\leq \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[\Phi(\widehat{u}(h^*_{\mathrm{PU}}), \widehat{v}(h^*_{\mathrm{PU}}), \widehat{p})\right] + \epsilon/2 \\ &\leq \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[\Phi(\widehat{u}(h^*_{\mathrm{ETU}}(\boldsymbol{x})), \widehat{v}(h^*_{\mathrm{ETU}}(\boldsymbol{x})), \widehat{p})\right] + \epsilon/2, \\ &\leq \Phi(u(h^*_{\mathrm{ETU}}(\boldsymbol{x})), v(h^*_{\mathrm{ETU}}(\boldsymbol{x})), p) + \epsilon,\end{aligned}$$

where we used Lemma 1 twice in the second and fourth inequality. Hence, with probability $1-\eta$,

$$\begin{aligned}\Big|\Phi(u(&h^*_{\mathrm{ETU}}(\boldsymbol{x})), v(h^*_{\mathrm{ETU}}(\boldsymbol{x})), p) \\ &- \Phi(u(h^*_{\mathrm{PU}}), v(h^*_{\mathrm{PU}}), p)\Big| \leq \epsilon.\end{aligned}$$

Using analogous argument, one can show that with probability $1-\delta$,

$$\begin{aligned}\Big|\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}&\left[\Phi(\widehat{u}(h^*_{\mathrm{ETU}}(\boldsymbol{x})), \widehat{v}(h^*_{\mathrm{ETU}}(\boldsymbol{x})), \widehat{p})\right] \\ &- \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[\Phi(\widehat{u}(h^*_{\mathrm{PU}}), \widehat{v}(h^*_{\mathrm{PU}}), \widehat{p})\right]\Big| \leq \epsilon,\end{aligned}$$

which finishes the proof.

B.3. Finite Sample Regime: Proof of Theorem 2

The PU-optimal classifier is:

$$h^*_{\mathrm{PU}} = \operatorname*{argmax}_h \Phi_{\mathrm{Prec}}(u(h), v(h), p) = \operatorname*{argmax}_h \frac{u(h)}{v(h)+\alpha}.$$

Proposition 2.

$$h^*_{\mathrm{PU}}(x) = \begin{cases} 1, \text{ if } x \in \mathcal{X}_1, \\ 0, \text{ else }. \end{cases}$$

Proof. Note that for the defined h^*_{PU} classifier, we have $u(h^*_{\mathrm{PU}}) = v(h^*_{\mathrm{PU}}) = \mathbb{P}(\mathcal{X}_1)$, and

$$\Phi_{\mathrm{Prec}}(u(h^*_{\mathrm{PU}}), v(h^*_{\mathrm{PU}}), p) = \frac{\mathbb{P}(\mathcal{X}_1)}{\mathbb{P}(\mathcal{X}_1)+\alpha}.$$

Firstly, observe that for any candidate optimal classifier h', it must hold that $h'(x) = 0$ for all $x \in \mathcal{X}_3$ (otherwise the metric strictly decreases). Now, suppose there exists a classifier $h' \neq h^*_{\mathrm{PU}}$ which has strictly higher utility than h^*_{PU}. Then, it must be that $h'(x) = 1$ for all

$x \in \mathcal{X}_2$. We have, $u(h') = \mathbb{P}(\mathcal{X}_1) + \mathbb{P}(\mathcal{X}_2)(1 - \sqrt{\alpha})$ and $v(h') = \mathbb{P}(\mathcal{X}_1) + \mathbb{P}(\mathcal{X}_2)$. So:

$$\Phi_{\text{Prec}}(u(h'), v(h'), p) = \frac{\mathbb{P}(\mathcal{X}_1) + \mathbb{P}(\mathcal{X}_2)(1 - \sqrt{\alpha})}{\mathbb{P}(\mathcal{X}_1) + \mathbb{P}(\mathcal{X}_2) + \alpha}.$$

But for the chosen small value of α, we can show the contradiction that:

$$\Phi_{\text{Prec}}(u(h'), v(h'), p) < \Phi_{\text{Prec}}(u(h^*_{\text{PU}}), v(h^*_{\text{PU}}), p).$$

Therefore, h^*_{PU} as stated is indeed optimal. □

We see from the above constructed example that the PU optimal classifier assigns negative labels to 50% of the data which are highly likely to belong to the positive class. PU is sensitive to label noise if the metric is less stable as implied by the high p-Lipschitz constant. Next, we show that ETU is relatively more robust.

Given a set of instances $\boldsymbol{x} = \{x_1, x_2, \dots, x_n\}$, recall that the ETU-optimal assignments can be computed as:

$$h^*_{\text{ETU}}(\boldsymbol{x}) = \mathbf{s}^* := \operatorname*{argmax}_{\mathbf{s} \in \{0,1\}^n} \mathbb{E}_{\boldsymbol{y} \sim \mathbb{P}(.|\boldsymbol{x})} \Phi_{\text{Prec}}(\mathbf{s}, \boldsymbol{y}) \; .$$

Proposition 3. *On the subset of instances in $\boldsymbol{x}$ that have deterministic labels, the* ETU*-optimal predictions satisfy:*

$$h^*_{\text{ETU}}(x_j) = s^*_j = \begin{cases} 1, & \text{if } x \in \mathcal{X}_1, \\ 0, & \text{if } x \in \mathcal{X}_3 \; . \end{cases}$$

Note that the predictions coincide with that of h^*_{PU} on these indices.

Proof. Let $\mathcal{I}_i = \{j : x_j \in \mathcal{X}_i\}$, for $i = 1, 2, 3$. Note that the optimal value at the solution $\mathbf{s}^*$ is given by:

$$\mathbb{E}_{\boldsymbol{y} \sim \mathbb{P}(.|\boldsymbol{x})} \Phi_{\text{Prec}}(\mathbf{s}^*, \boldsymbol{y}) = \frac{\sum_{j \in \mathcal{I}_1} s^*_j + \Delta(\mathbf{s}^*_{\mathcal{I}_2}, \boldsymbol{y}_{\mathcal{I}_2})}{\sum_{j \in \mathcal{I}_1 \cup \mathcal{I}_3} s^*_j + \sum_{j \in \mathcal{I}_2} s^*_j + \alpha n}, \tag{2}$$

where $\mathbf{s}^*_{\mathcal{I}_2}$ indicates the optimal assignments corresponding to indices in $\mathcal{I}_2$ and $\Delta(\mathbf{s}^*_{\mathcal{I}_2}, \boldsymbol{y}_{\mathcal{I}_2})$ is a quantity that depends only on indices in $\mathcal{I}_2$, and is given by:

$$\Delta(\mathbf{s}^*_{\mathcal{I}_2}, \boldsymbol{y}_{\mathcal{I}_2}) = \sum_{\mathbf{y}_{\mathcal{I}_2} \in \{0,1\}^{|\mathcal{I}_2|}} \mathbb{P}(\mathbf{y}_{\mathcal{I}_2}) \langle \mathbf{y}_{\mathcal{I}_2}, \mathbf{s}^*_{\mathcal{I}_2} \rangle \tag{3}$$

Fixing the optimal predictions for indices corresponding to $\mathcal{I}_2$, the value (2) is maximized by maximizing the numerator term $\sum_{j \in \mathcal{I}_1} s^*_j$ and minimizing the denominator term $\sum_{j \in \mathcal{I}_1 \cup \mathcal{I}_3} s^*_j$. This is achieved precisely when the optimal solution satisfies the statement in the proposition. The proof is complete. □

We know from Proposition 2 that h^*_{PU} sets the labels corresponding to indices in the set $\mathcal{I}_2$ to 0. Now let us examine what happens in the case of ETU, when labels have mild noise (i.e. with some small probability $\sqrt{\epsilon}$, the label of an instance from $\mathcal{X}_2$ can be 0), at optimality. Consider a candidate optimal solution $\mathbf{s}'$ that behaves exactly like h^*_{PU}, i.e. $\mathbf{s}'_j = 0$ for all $j \in \mathcal{I}_2$, for some $1 \le k \le |\mathcal{I}_2|$.

Then, $\Delta(\mathbf{s}'_{\mathcal{I}_2}, \boldsymbol{y}_{\mathcal{I}_2}) = 0$, so:

$$\mathbb{E}_{\boldsymbol{y} \sim \mathbb{P}(.|\boldsymbol{x})} \Phi_{\text{Prec}}(\mathbf{s}', \boldsymbol{y}) = \frac{|\mathcal{I}_1|}{|\mathcal{I}_1| + \alpha n} \; . \tag{4}$$

Now, consider another candidate solution $\mathbf{s}''$ that is equal to $\mathbf{s}'$, but has a value of 1 corresponding to a subset of indices $j_1, j_2, \dots, j_k \in \mathcal{I}_2$. The value of this solution can be shown to be:

$$\mathbb{E}_{\boldsymbol{y} \sim \mathbb{P}(.|\boldsymbol{x})} \Phi_{\text{Prec}}(\mathbf{s}'', \boldsymbol{y}) = \frac{|\mathcal{I}_1| + k(1-\epsilon)}{|\mathcal{I}_1| + k + \alpha n} \; . \tag{5}$$

Comparing equations (4) and (5), we have that if:

$$\epsilon < \frac{\alpha n}{|\mathcal{I}_1| + \alpha n}, \tag{6}$$

then $\mathbf{s}''$ is a strictly better solution than $\mathbf{s}'$. In particular, as (5) is mononotic in k, the optimal choice is $k = |\mathcal{I}_2|$. This immediately leads to the following corollary.

Corollary 1. *1. If $|\mathcal{I}_2| = 0$, then*

$$h^*_{\text{ETU}}(\boldsymbol{x}) := \mathbf{s}^* = h^*_{\text{PU}}(\boldsymbol{x}) \; .$$

2. Otherwise, if $\epsilon < \frac{\alpha}{1+\alpha}$, then

$$h^*_{\text{ETU}}(\boldsymbol{x}) := \mathbf{s}^* \neq h^*_{\text{PU}}(\boldsymbol{x}) \; .$$

In particular, h^*_{ETU} assigns label 1 to all instances that are overwhelmingly positive under $\mathbb{P}$, corresponding to indices $\mathcal{I}_2$, whereas h^*_{PU} assigns label 0.

3. If $|\mathcal{I}_1| = 0$, but $|\mathcal{I}_2| > 0$ then for any $0 < \epsilon < 1$,

$$h^*_{\text{ETU}}(\boldsymbol{x}) := \mathbf{s}^* \neq h^*_{\text{PU}}(\boldsymbol{x}) := \mathbf{0} \; .$$

Note that $\epsilon < \alpha/(1+\alpha)$ does *not* hold for our choice of $\epsilon = \sqrt{\alpha}$. However, case 3 in Corollary 1 is sufficient to establish the bound in Theorem 2, when $\mathbb{P}(\mathcal{X}_2)$ is very large.

C. Proofs for Section 4.1

Fix a binary classifier $h \colon X \to \{0, 1\}$ and let the input sample $\boldsymbol{x} = (x_1, \dots, x_n)$ be generated i.i.d. from $\mathbb{P}$. For the sake of clarity, abbreviate $\eta(x_i) = \eta_i$ and $h(x_i) = h_i$, $i = 1, \dots, n$. In the proofs of Lemma 2 and Lemma 3 we will use the following:

- *Empirical quantities*:

$$\widehat{u}(h) = \frac{1}{n}\sum_{i=1}^{n} h_i y_i, \widehat{v}(h) = \frac{1}{n}\sum_{i=1}^{n} h_i, \widehat{p} = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

- *Semi-empirical quantities*:

$$\widetilde{u}(h) = \frac{1}{n}\sum_{i=1}^{n} h_i \eta_i, \qquad \text{and} \quad \widetilde{p} = \frac{1}{n}\sum_{i=1}^{n} \eta_i$$

(we do not define $\widetilde{v}(h)$, as it would the same as $\widehat{v}(h)$).

Note that:

$$\widetilde{u}(h) = \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[\widehat{u}(h)\right], \qquad \text{and} \quad \widetilde{p} = \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[\widehat{p}\right].$$

We will jointly denote $\widehat{\boldsymbol{z}} = (\widehat{u}(h), \widehat{p})$, and similarly $\widetilde{\boldsymbol{z}} = (\widetilde{u}(h), \widetilde{p})$. We will also abbreviate $\Phi(\widehat{\boldsymbol{z}}) = \Phi(\widehat{u}(h), \widehat{v}(h), \widehat{p})$ and similarly for $\Phi(\widetilde{\boldsymbol{z}})$.

C.1. Proof of Lemma 2

Assume Φ is two-times differentiable, with all partial second-order derivatives bounded by A. Taylor expanding $\Phi(\widehat{\boldsymbol{z}})$ around point $\widetilde{\boldsymbol{z}}$ up to the second order gives:

$$\begin{aligned}\Phi(\widehat{\boldsymbol{z}}) &= \Phi(\widetilde{\boldsymbol{z}}) + \nabla\Phi(\widetilde{\boldsymbol{z}})^\top(\widehat{\boldsymbol{z}} - \widetilde{\boldsymbol{z}}) \\ &+ \frac{1}{2}(\widehat{\boldsymbol{z}} - \widetilde{\boldsymbol{z}})^\top \nabla^2\Phi(\boldsymbol{z})(\widehat{\boldsymbol{z}} - \widetilde{\boldsymbol{z}})\end{aligned}$$

for some $\boldsymbol{z}$ between $\widehat{\boldsymbol{z}}$ and $\widetilde{\boldsymbol{z}}$. Note that $\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[\widehat{\boldsymbol{z}}\right] = \widetilde{\boldsymbol{z}}$, so that:

$$\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[\nabla\Phi(\widetilde{\boldsymbol{z}})^\top(\widehat{\boldsymbol{z}} - \widetilde{\boldsymbol{z}})\right] = 0.$$

Furthermore, note that:

$$\begin{aligned}&(\widehat{\boldsymbol{z}} - \widetilde{\boldsymbol{z}})^\top \nabla^2\Phi(\boldsymbol{z})(\widehat{\boldsymbol{z}} - \widetilde{\boldsymbol{z}}) \\ &\quad = \nabla^2_{uu}(\widehat{u} - \widetilde{u})^2 + 2\nabla^2_{up}(\widehat{u} - \widetilde{u})(\widehat{p} - \widetilde{p}) + \nabla^2_{pp}(\widehat{p} - \widetilde{p})^2 \\ &\quad \le A\big((\widehat{u} - \widetilde{u})^2 + 2|(\widehat{u} - \widetilde{u})(\widehat{p} - \widetilde{p})| + (\widehat{p} - \widetilde{p})^2\big) \\ &\quad \le 2A\big((\widehat{u} - \widetilde{u})^2 + (\widehat{p} - \widetilde{p})^2\big),\end{aligned}$$

where we used elementary inequality $ab \le a^2 + b^2$, and $\nabla^2_{uu}, \nabla^2_{up}, \nabla^2_{pp}$ denote the second-order derivatives evaluated at some $\boldsymbol{z} = (u, p)$. Hence:

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[(\widehat{\boldsymbol{z}} - \widetilde{\boldsymbol{z}})^\top \nabla^2\Phi(\widetilde{\boldsymbol{z}})(\widehat{\boldsymbol{z}} - \widetilde{\boldsymbol{z}})\right] \\ &\qquad \le 2A\left(\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[(\widehat{u} - \widetilde{u})^2\right] + \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[(\widehat{p} - \widetilde{p})^2\right]\right).\end{aligned}$$

Since $\widehat{u}$ is the empirical average over n labels and $\widetilde{u}$ is its expectation (over the labels), $\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[(\widehat{u} - \widetilde{u})^2\right]$ is the variance of $\widehat{u}$, which is at most $\frac{1}{4n}$, because $\widehat{u} \in [0, 1]$:

$$\mathrm{var}(\widehat{u}) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{var}(h_i y_i) \le \frac{1}{n}\sum_{i=1}^{n} h_i \eta_i (1 - \eta_i) \le \frac{1}{4n},$$

where we used the independence of labels y_i, $i = 1, \ldots, n$. Similarly, $\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[(\widehat{p} - \widetilde{p})^2\right]$ is at most $\frac{1}{4n}$, which in total gives:

$$\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[(\widehat{\boldsymbol{z}} - \widetilde{\boldsymbol{z}})^\top \nabla^2\Phi(\widetilde{\boldsymbol{z}})(\widehat{\boldsymbol{z}} - \widetilde{\boldsymbol{z}})\right] \le \frac{A}{n}.$$

Using a lower bound $-A$ on the second-order derivatives and performing a similar chain of reasoning, one also gets:

$$\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[(\widehat{\boldsymbol{z}} - \widetilde{\boldsymbol{z}})^\top \nabla^2\Phi(\widetilde{\boldsymbol{z}})(\widehat{\boldsymbol{z}} - \widetilde{\boldsymbol{z}})\right] \ge -\frac{A}{n}.$$

From that we have:

$$\|\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[\Phi(\widehat{\boldsymbol{z}})\right] - \Phi(\widetilde{\boldsymbol{z}})\| \le \frac{A}{2n},$$

which is exactly what was to be shown.

C.2. Proof of Lemma 3

Assume Φ is three-times differentiable, with all partial third-order derivatives bounded by B. Taylor expanding $\Phi(\widehat{\boldsymbol{z}})$ around point $\widetilde{\boldsymbol{z}}$ up to the third order gives:

$$\begin{aligned}\Phi(\widehat{\boldsymbol{z}}) &= \Phi(\widetilde{\boldsymbol{z}}) + \nabla\Phi(\widetilde{\boldsymbol{z}})^\top(\widehat{\boldsymbol{z}} - \widetilde{\boldsymbol{z}}) \\ &+ \frac{1}{2}(\widehat{\boldsymbol{z}} - \widetilde{\boldsymbol{z}})^\top \nabla^2\Phi(\widetilde{\boldsymbol{z}})(\widehat{\boldsymbol{z}} - \widetilde{\boldsymbol{z}}) \\ &+ \frac{1}{6}\sum_{\alpha,\beta,\gamma=1}^{2} \frac{\partial^3\Phi(\boldsymbol{z})}{\partial z_\alpha \partial z_\beta \partial z_\gamma}(\widehat{z}_\alpha - \widetilde{z}_\alpha)(\widehat{z}_\beta - \widetilde{z}_\beta)(\widehat{z}_\gamma - \widetilde{z}_\gamma),\end{aligned}$$

for some $\boldsymbol{z}$ between $\widehat{\boldsymbol{z}}$ and $\widetilde{\boldsymbol{z}}$. First note that $\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[\widehat{\boldsymbol{z}}\right] = \widetilde{\boldsymbol{z}}$, so that:

$$\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[\nabla\Phi(\widetilde{\boldsymbol{z}})^\top(\widehat{\boldsymbol{z}} - \widetilde{\boldsymbol{z}})\right] = 0.$$

Furthermore,

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[\nabla^2(\widehat{\boldsymbol{z}} - \widetilde{\boldsymbol{z}})^\top \Phi(\widetilde{\boldsymbol{z}})(\widehat{\boldsymbol{z}} - \widetilde{\boldsymbol{z}})\right] \\ &\qquad = \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[\mathrm{tr}\left(\nabla^2\Phi(\widetilde{\boldsymbol{z}})(\widehat{\boldsymbol{z}} - \widetilde{\boldsymbol{z}})(\widehat{\boldsymbol{z}} - \widetilde{\boldsymbol{z}})^\top\right)\right] \\ &\qquad = \mathrm{tr}\left(\nabla^2\Phi(\widetilde{\boldsymbol{z}})\Sigma\right),\end{aligned}$$

where $\Sigma = \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[(\widehat{\boldsymbol{z}} - \widetilde{\boldsymbol{z}})(\widehat{\boldsymbol{z}} - \widetilde{\boldsymbol{z}})^\top\right]$ is the covariance matrix of $\widehat{\boldsymbol{z}} - \widetilde{\boldsymbol{z}}$. By independence of examples,

$$\begin{aligned}\Sigma &= \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{E}_{y_i|x_i}\left[\begin{pmatrix} h_i(y_i - \eta_i)^2 & h_i(y_i - \eta_i)^2 \\ h_i(y_i - \eta_i)^2 & (y_i - \eta_i)^2 \end{pmatrix}\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n} \eta_i(1 - \eta_i)\begin{pmatrix} h_i & h_i \\ h_i & 1 \end{pmatrix},\end{aligned}$$

so that:

$$\mathrm{tr}\left(\nabla^2\Phi(\widetilde{\boldsymbol{z}})\Sigma\right) = (\nabla^2_{uu} + 2\nabla^2_{up})s_u + \nabla^2_{pp}s_p,$$

where:

$$s_p := \frac{1}{n^2}\sum_{i=1}^n \eta_i(1-\eta_i),$$
$$s_u := \frac{1}{n^2}\sum_{i=1}^n h_i\eta_i(1-\eta_i),$$

and $\nabla^2_{uu}, \nabla^2_{up}, \nabla^2_{pp}$ denote be the second-order derivative terms evaluated at $(\widetilde{u}, \widetilde{p})$. Thus, to finish the proof, we only need to show that the first order term is bounded by $\frac{B}{3}n^{-3/2}$. To this end, note that for any numbers a_i, b_{ijk}, such that $|b_{ijk}| \le B$, $i,j,k = 1,\ldots,2$:

$$\sum_{ijk} b_{ijk}a_ia_ja_k \le B\sum_{ijk}|a_i||a_j||a_k| = B(|a_1|+|a_2|)^3.$$

By Hölder's inequality,

$$\sum_{i=1}^2 |a_i| \le \left(\sum_{i=1}^2 |a_i|^3\right)^{1/3} 2^{2/3},$$

so that:

$$B(|a_1|+|a_2|+|a_3|)^3 \le 4B\left(|a_1|^3+|a_2|^3+|a_3|^3\right).$$

Hence, if we bound:

$$\frac{\partial^3\Phi(\boldsymbol{z})}{\partial z_\alpha \partial z_\beta \partial z_\gamma} \le B,$$

the third-order term $\frac{1}{6}\sum_{\alpha,\beta,\gamma=1}^2 \ldots$ is bounded by:

$$\frac{2B}{3}\left(|\widehat{u}-\widetilde{u}|^3 + |\widehat{p}-\widetilde{p}|^3\right)$$

We now bound $\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[|\widehat{u}-\widetilde{u}|^3\right]$ and $\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[|\widehat{p}-\widetilde{p}|^3\right]$. By Cauchy-Schwarz inequality,

$$\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[|\widehat{p}-\widetilde{p}|^3\right] \le \sqrt{\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[(\widehat{p}-\widetilde{p})^4\right]}\sqrt{\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[(\widehat{p}-\widetilde{p})^2\right]}.$$

Before, we already showed that

$$\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[(\widehat{p}-\widetilde{p})^2\right] \le \frac{1}{4n}.$$

Denote $a_i = y_i - \eta_i$, and let $\mu_k = \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[a_i^k\right]$. Using $\mu_1 = 0$, we have:

$$\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[(\widehat{p}-\widetilde{p})^4\right] = \frac{1}{n^4}\sum_{i,j,k,\ell} a_ia_ja_ka_\ell$$
$$= \frac{1}{n^4}\left(n\mu_4 + 3n(n-1)\mu_2^2\right).$$

Since $\mu_2 \le \frac{1}{4}$ and $\mu_4 \le \frac{1}{12}$, $\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[(\widehat{p}-\widetilde{p})^4\right] \le \frac{3}{16n^2}$, and thus:

$$\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[|\widehat{p}-\widetilde{p}|^3\right] \le \frac{\sqrt{3}}{8}n^{-3/2} \le \frac{1}{4}n^{-3/2}.$$

Using similar bound for $\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[|\widehat{u}-\widetilde{u}|^3\right]$, we conclude that the third-order term is bounded by $\frac{B}{3}n^{-3/2}$. Bounding the third-order derivatives from below by $-B$, and using similar reasoning gives a lower bound of the same value. This finishes the proof.

C.3. Proof of Theorem 3

Abbreviating $\Phi(h) = \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[\Phi(\widehat{u}(h), \widehat{v}(h), \widehat{p})\right]$ and $\Phi_a(h) = \Phi_{\text{appr}}(h)$:

$$\Phi(h^*_{\text{ETU}}) - \Phi(h^*_a) = \underbrace{\Phi(h^*_{\text{ETU}}) - \Phi_a(h^*_{\text{ETU}})}_{\le \frac{B}{3n^{3/2}}}$$
$$\underbrace{\Phi_a(h^*_{\text{ETU}}) - \Phi_a(h^*_a)}_{\le 0} + \underbrace{\Phi_a(h^*_a) - \Phi(h^*_a)}_{\le \frac{B}{3n^{3/2}}} \le \frac{2B}{3n^{3/2}},$$

where the bounds shown in the inequalities are from Lemma 3.

C.4. Derivation of the approximation algorithm for F_β-measure

Recall that $F_\beta(u,v,p) = \frac{(1+\beta^2)u}{\beta^2 p + v}$. The seconder order derivatives with respect to u and p are:

$$\frac{\partial^2 F_\beta}{\partial u^2} = 0, \frac{\partial^2 F_\beta}{\partial u \partial p} = \frac{-\beta^2(1+\beta^2)}{(\beta^2 p + v)^2}, \frac{\partial^2 F_\beta}{\partial p^2} = \frac{2\beta^4(1+\beta^2)u}{(\beta^2 p + v)^3}.$$

To optimize $\Phi_{\text{appr}}(h)$, we first sort observations according to $\eta(x_i)$. Then we precompute:

$$\widetilde{p} = \frac{1}{n}\sum_{i=1}^n \eta(x_i), \qquad \widetilde{p}_{\text{var}} = \frac{1}{n^2}\sum_{i=1}^n \eta(x_i)(1-\eta(x_i)).$$

Next, for each $k = 0, 1, \ldots, n$, we precompute:

$$\widetilde{u}^k = \frac{1}{n}\sum_{i=1}^k \eta(x_i), \widehat{v}^k = \frac{k}{n}, \widetilde{u}^k_{\text{var}} = \frac{1}{n^2}\sum_{i=1}^k \eta(x_i)(1-\eta(x_i)).$$

We then choose k for which the ETU approximation:

$$\frac{(1+\beta^2)\widetilde{u}^k}{\beta^2\widetilde{p} + \frac{k}{n}} - \frac{\beta^2(1+\beta^2)}{(\beta^2\widetilde{p} + \frac{k}{n})^2}\widetilde{u}^k_{\text{var}} + \frac{\beta^4(1+\beta^2)\widetilde{u}^k}{(\beta^2\widetilde{p} + \frac{k}{n})^3}\widetilde{p}_{\text{var}},$$

is maximized. The maximization can be done in time linear in $O(n)$, so the most expensive operation is sorting the instances.

D. Additional material to Section 4.2

Let $\boldsymbol{x} = (x_1, \ldots, x_n)$ be the input sample (test set) of size n generated i.i.d. from $\mathbb{P}$. Given $\boldsymbol{x}$ and a function $\widehat{\eta}\colon X \to [0,1]$, let

$$\widehat{h} = \operatorname*{argmax}_{h \in \widehat{\mathcal{H}}} \underbrace{\mathbb{E}_{\boldsymbol{y}\sim\widehat{\eta}(\boldsymbol{x})}\left[\Phi(\widehat{u}(h), \widehat{v}(h), \widehat{p})\right]}_{=:\widehat{\Phi}_{\text{ETU}}(h)}.$$

be the classifier returned by the ETU procedure upon receiving the input sample $\boldsymbol{x}$. Likewise, let:

$$h^* = \operatorname*{argmax}_{h \in \widehat{\mathcal{H}}} \underbrace{\mathbb{E}_{\boldsymbol{y} \sim \eta(\boldsymbol{x})} \left[\Phi(\widehat{u}(h), \widehat{v}(h), \widehat{p})\right]}_{=:\Phi_{\mathrm{ETU}}(h)},$$

be the optimal ETU classifier in $\widehat{\mathcal{H}}$. We want to bound the difference $\mathbb{E}_{\boldsymbol{x}}\left[\left|\Phi_{\mathrm{ETU}}(\widehat{h}) - \Phi_{\mathrm{ETU}}(h^*)\right|\right]$. By the definition of h^*, $\Phi_{\mathrm{ETU}}(\widehat{h}) \leq \Phi_{\mathrm{ETU}}(h^*)$ for any $\boldsymbol{x}$, and thus:

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}}\left[\left|\Phi_{\mathrm{ETU}}(\widehat{h}) - \Phi_{\mathrm{ETU}}(h^*)\right|\right] \\
&\quad = \mathbb{E}_{\boldsymbol{x}}\left[\Phi_{\mathrm{ETU}}(h^*)\right] - \mathbb{E}_{\boldsymbol{x}}\left[\Phi_{\mathrm{ETU}}(\widehat{h})\right] \\
&\quad = \mathbb{E}_{\boldsymbol{x}}\left[\Phi_{\mathrm{ETU}}(h^*)\right] - \mathbb{E}_{\boldsymbol{x}}\left[\widehat{\Phi}_{\mathrm{ETU}}(h^*)\right] \\
&\qquad + \underbrace{\mathbb{E}_{\boldsymbol{x}}\left[\widehat{\Phi}_{\mathrm{ETU}}(h^*)\right] - \mathbb{E}_{\boldsymbol{x}}\left[\widehat{\Phi}_{\mathrm{ETU}}(\widehat{h})\right]}_{\leq 0} \\
&\qquad + \mathbb{E}_{\boldsymbol{x}}\left[\widehat{\Phi}_{\mathrm{ETU}}(\widehat{h})\right] - \mathbb{E}_{\boldsymbol{x}}\left[\Phi_{\mathrm{ETU}}(\widehat{h})\right] \\
&\quad \leq 2 \sup_{h \in \widehat{\mathcal{H}}} \left|\mathbb{E}_{\boldsymbol{x}}\left[\Phi_{\mathrm{ETU}}(h) - \widehat{\Phi}_{\mathrm{ETU}}(h)\right]\right|. \qquad (7)
\end{aligned}$$

Now, fix some classifier h and input sample $\boldsymbol{x}$. We let $\widehat{u}(h), \widehat{v}(h), \widehat{p}$ denote the random variables generated according to η (for fixed $\boldsymbol{x}$), while $\widehat{u}'(h), \widehat{p}'(h)$ denote random variables generated according to $\widehat{\eta}$; for instance, $\widehat{u}'(h) = \frac{1}{n}\sum_{i=1}^n h(x_i) y_i$, where $y_i \sim \widehat{\eta}(x_i)$. Using this notation, we have:

$$\begin{aligned}
\Phi_{\mathrm{ETU}}(h) &= \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[\Phi(\widehat{u}(h), \widehat{v}(h), \widehat{p})\right], \\
\widehat{\Phi}_{\mathrm{ETU}}(h) &= \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[\Phi(\widehat{u}'(h), \widehat{v}(h), \widehat{p}')\right]
\end{aligned}$$

(note that $\widehat{v}(h)$ does not depend on $\widehat{\eta}$ or η, we $\widehat{v}'(h) = \widehat{v}(h)$). We now bound the term under sup in (7):

$$\begin{aligned}
&\left|\mathbb{E}_{\boldsymbol{x}}\left[\Phi_{\mathrm{ETU}}(h) - \widehat{\Phi}_{\mathrm{ETU}}(h)\right]\right| \\
&\quad \leq \mathbb{E}\left[\left|\Phi(\widehat{u}, \widehat{v}, \widehat{p}) - \Phi(\widehat{u}', \widehat{v}, \widehat{p}')\right|\right] \\
&\quad \leq \mathbb{E}\left[\left|\Phi(\widehat{u}, \widehat{v}, \widehat{p}) - \Phi(u, v, p)\right|\right] \\
&\qquad + \mathbb{E}\left[\left|\Phi(u, v, p) - \Phi(\widehat{u}', \widehat{v}, \widehat{p}')\right|\right],
\end{aligned}$$

where the first inequality is due to Jensen's inequality applied to a convex function $x \mapsto |x|$, the all expectations except for the first line are joint with respect to $(\boldsymbol{x}, \boldsymbol{y})$, and for the sake of clarity we drop the dependence on h in $\widehat{u}(h), \widehat{v}(h), \widehat{u}'(h)$. Now, it follow from Lemma 1 that:

$$\mathbb{E}\left[\left|\Phi(\widehat{u}, \widehat{v}, \widehat{p}) - \Phi(u, v, p)\right|\right] \leq c\sqrt{\frac{\log n}{n}},$$

for some constant c. Moreover, using p-Lipschitzness of Φ, we have:

$$\begin{aligned}
\mathbb{E}\left[\left|\Phi(u, v, p) - \Phi(\widehat{u}', \widehat{v}, \widehat{p}')\right|\right] &\leq U_p \mathbb{E}\left[|\widehat{u}' - u|\right] \\
&\quad + V_p \mathbb{E}\left[|\widehat{v} - v|\right] + P_p \mathbb{E}\left[|\widehat{p}' - p|\right].
\end{aligned}$$

Now, the term $\mathbb{E}\left[|\widehat{v} - v|\right]$ is well-controlled and was shown in the proof of Lemma 1 to be at most $\sqrt{\frac{1}{4n}}$ as the expected deviation of the empirical average of $[0,1]$-valued random variable from its mean. Thus it remains to bound the terms $\mathbb{E}\left[|\widehat{p}' - p|\right]$ and $\mathbb{E}\left[|\widehat{u}' - u|\right]$. Define:

$$\begin{aligned}
\widetilde{p}' &= \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[\widehat{p}'\right] = \frac{1}{n}\sum_{i=1}^n \widehat{\eta}(x_i), \\
\widetilde{u}' &= \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[\widehat{u}'\right] = \frac{1}{n}\sum_{i=1}^n h(x_i)\widehat{\eta}(x_i), \\
p_{\widehat{\eta}} &= \mathbb{E}_{\boldsymbol{x}}\left[\widetilde{p}'\right] = \mathbb{E}\left[\widehat{\eta}(x)\right]. \\
u_{\widehat{\eta}} &= \mathbb{E}_{\boldsymbol{x}}\left[\widetilde{u}'\right] = \mathbb{E}\left[h(x)\widehat{\eta}(x)\right].
\end{aligned}$$

We decompose:

$$|p - \widehat{p}'| \leq |p - p_{\widehat{\eta}}| + |p_{\widehat{\eta}} - \widetilde{p}'| + |\widetilde{p}' - \widehat{p}'|$$

As before, we use the fact that $\mathbb{E}_{\boldsymbol{x}}\left[|p_{\widehat{\eta}} - \widetilde{p}'|\right]$, as well as $\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[|\widetilde{p}' - \widehat{p}'|\right]$ are both the expected deviations of the empirical averages of $[0,1]$-valued random variables from their means, and therefore are bounded by $\sqrt{\frac{1}{4n}}$. Hence:

$$\mathbb{E}\left[|\widehat{p}' - p|\right] \leq |p - p_{\widehat{\eta}}| + \frac{1}{\sqrt{n}}.$$

Using analogous way of reasoning, one gets:

$$\mathbb{E}\left[|\widehat{u}' - u|\right] \leq |u - u_{\widehat{\eta}}| + \frac{1}{\sqrt{n}}.$$

Putting it all together, we get:

$$\begin{aligned}
&\left|\mathbb{E}_{\boldsymbol{x}}\left[\Phi_{\mathrm{ETU}}(h) - \widehat{\Phi}_{\mathrm{ETU}}(h)\right]\right| \\
&\quad \leq c'\sqrt{\frac{\log n}{n}} + U_p|u(h) - u_{\widehat{\eta}}(h)| + P_p|p - p_{\widehat{\eta}}|,
\end{aligned}$$

for some constant c'. Using (7), we finally get:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{x}}\left[\left|\Phi_{\mathrm{ETU}}(\widehat{h}) - \Phi_{\mathrm{ETU}}(h^*)\right|\right] &\leq c'\sqrt{\frac{\log n}{n}} + P_p|p - p_{\widehat{\eta}}| \\
&\quad + \sup_{h \in \widehat{\mathcal{H}}} U_p|u(h) - u_{\widehat{\eta}}(h)|,
\end{aligned}$$

which was to be shown.

E. Isotron Algorithm (Kalai & Sastry, 2009)

Here we include the Isotron Algorithm of (Kalai & Sastry, 2009) for completeness. The second update step is the Pool of Adjacent Violators (PAV) routine, which solves the isotonic regression problem:

$$u_1^*, u_2^*, \ldots, u_n^* = \arg \min_{u_1 \leq u_2 \leq \cdots \leq u_n} \sum_{i=1}^{n} (y_i - u_i)^2,$$

where the instances are assumed to be sorted according to their scores $\mathbf{w}^T x$ (using $\mathbf{w}$ obtained in first update step of the iteration). This is a convex quadratic program and can be solved efficiently. The output link function u of the Algorithm is a piecewise linear estimate.

Algorithm 2 The Isotron algorithm (Kalai & Sastry, 2009).

Input: Training data $\{(x_i, y_i)\}_{i=1}^n$, iterations T
Output: $\mathbf{w}_T, u_T$
$\mathbf{w}_0 \leftarrow 0$
$u_0 \leftarrow z \mapsto \min(\max(0, 2 \cdot z + 1), 1)$
for $t = 1, 2, \ldots, T$ **do**
  $\mathbf{w}_t \leftarrow \mathbf{w}_{t-1} + \frac{1}{n} \sum_{i=1}^n (y_i - u_{t-1}(\langle \mathbf{w}_{t-1}, x_i \rangle)) \cdot x_i$
  $u_t \leftarrow \text{PAV}(\{\langle \mathbf{w}_t, x_i \rangle, y_i\})$
end for