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 $\Phi$ amounts to showing that:

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

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, then $\Phi(u, v, p) \geq G_p$ for some positive constant $G_p$ (which may depend on $p$), and $|\Phi(u, v, p)| \leq \Phi_{\max}$ 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 $B, B'$:

$$\Phi - \Phi' = \frac{A - \frac{A'}{B'} B}{B} = \frac{A - A' + \frac{A'}{B'} B' - \frac{A'}{B'} B}{B},$$

hence:

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

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

$$|\Phi - \Phi'| \leq 2\Delta u - \Delta v - \Delta p,$$

so that by triangle inequality:

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

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:

$$A - A' = u - vp - u' + v'p' = \Delta u + (vp' - vp) + (v'p - vp') = \Delta u - v\Delta p - p'\Delta v,$$

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

$$B - B' = 2p(1-p) - 2p'(1-p') = 2(p - p') + 2(p'^2 - p^2) = 2(1 - p' - p)(p - p'),$$

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

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 $A$:

$$A - A' = (1 - v - p + u)(u - u') + u'(u - p - v - u' - p' - v') = (1 - v - p + u)(\Delta u + u'(\Delta u - \Delta v - \Delta p)), $$

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

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 $A$:

$$A - A' = (v - u)(p - u') - (v' - u')(p - u) + (v' - u')(p - u') - (v' - u')(p' - u') = (\Delta v - \Delta u)(p - u) + (v' - u')(\Delta p - \Delta u),$$

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

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 $u, v, \hat{u}, \hat{v}$ to denote $u(h), v(h), \hat{u}(h), \hat{v}(h)$. Due to the Lipschitz assumption:

$$|\Phi(u, v, p) - \Phi(\hat{u}, \hat{v}, \hat{p})| \leq U_p|u - \hat{u}| + V_p|v - \hat{v}| + P_p|p - \hat{p}|.$$
Denote:

\[ p = E_{y|x}[\hat{p}] = \frac{1}{n} \sum_{i=1}^{n} \eta(x_i), \]

\[ \tilde{u} = E_{y|x}[\tilde{u}] = \frac{1}{n} \sum_{i=1}^{n} h(x_i) \eta(x_i) \]

We have:

\[ E_{y|x}[|p - \tilde{p}|] = E_{y|x}[|p - \tilde{p}| + \tilde{p} - \tilde{p}|] \]

\[ \leq |p - \tilde{p}| + E_{y|x}[(\tilde{p} - \tilde{p})] \]

\[ = |p - \tilde{p}| + \sqrt{E_{y|x}[(\tilde{p} - \tilde{p})^2]} \]

\[ \leq |p - \tilde{p}| + \sqrt{\text{Var}_{y|x}(\tilde{p})} \leq |p - \tilde{p}| + \frac{1}{\sqrt{4n}}, \]

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:

\[ E_{y|x}[|u - \tilde{u}|] \leq |u - \tilde{u}| + \sqrt{\frac{u}{4n}} \leq |u - \tilde{u}| + \frac{1}{\sqrt{4n}}. \]

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

\[ \left| \Phi(u, v, p) - E_{y|x}[\Phi(\tilde{u}, \tilde{v}, \tilde{p})] \right| \leq E_{y|x}[|\Phi(u, v, p) - \Phi(\tilde{u}, \tilde{v}, \tilde{p})|], \]

so that:

\[ \left| \Phi(u, v, p) - E_{y|x}[\Phi(\tilde{u}, \tilde{v}, \tilde{p})] \right| \leq U_p |u - \tilde{u}| + V_p |v - \tilde{v}| + P_p |p - \tilde{p}| + \frac{U_p + V_p}{2\sqrt{n}}. \]

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 \( 1 - \delta/2 \),

\[ |p - \tilde{p}| \leq \sqrt{\frac{\log \frac{1}{\delta}}{2n}}. \]

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

\[ |v - \tilde{v}| \leq 2E_{\xi}|\mathcal{R}_{\xi}(\mathcal{H})| + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}, \]

and similarly, with probability \( 1 - \delta/4 \),

\[ |u - \tilde{u}| \leq 2E_{\xi}|\mathcal{R}_{\xi}(\mathcal{H}_{\eta})| + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}, \]

where \( \mathcal{H}_{\eta} = \{ h \cdot \eta : h \in \mathcal{H} \} \), and:

\[ \mathcal{R}_{\xi}(\mathcal{H}) = E_{\xi}\left[ \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i h(x_i) \right] \]

is the Rademacher complexity 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 \( \mathbb{E}[z_i] = \eta(x_i) \), we have:

\[ \sum_{i=1}^{n} \sigma_i h(x_i) \eta(x_i) = \mathbb{E}_x \left[ \sum_{i=1}^{n} \sigma_i h(x_i) z_i \right], \]

so that:

\[ \mathcal{R}_{\xi}(\mathcal{H}_{\eta}) = E_{\xi}\left[ \sup_{h \in \mathcal{H}/n} \frac{1}{n} \mathbb{E}_x \left[ \sum_{i=1}^{n} \sigma_i h(x_i) z_i \right] \right] \]

\[ \leq E_{\xi, z}\left[ \sup_{h \in \mathcal{H}/n} \frac{1}{n} \sum_{i=1}^{n} \sigma_i h(x_i) z_i \right] \]

\[ = E_{\xi}\left[ \sup_{h \in \mathcal{H}/n} \frac{1}{n} \sum_{i=1}^{n} \sigma_i h(x_i) \right] = \mathcal{R}_{\xi}(H), \]

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

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

\[ \left| \Phi(u, v, p) - E_{y|x}[\Phi(\tilde{u}, \tilde{v}, \tilde{p})] \right| \leq 4L_p E_{\xi}|\mathcal{R}_{\xi}(\mathcal{H})| \]

\[ + 3L_p \sqrt{\frac{\log \frac{1}{\delta}}{2n}} + \frac{L_p}{\sqrt{n}}. \]

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

\[ \mathcal{R}_{\xi}(\mathcal{H}) \leq \sqrt{\frac{2\log(n+1)}{n}}. \]

\(^6\) Variables \( \sigma_i, i = 1, \ldots, n \), 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 $\tilde{u}(h) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} y_i$. Hence, $\left| \Phi(u, v, p) - \mathbb{E}_{y|x} \Phi(\tilde{u}, \tilde{v}, \tilde{p}) \right| = |p - \tilde{p}|$, where $\tilde{p} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \eta(x_i)$ and $\mathbb{E}_{x}[\tilde{p}] = p$. Assume that $\eta(x)$ follows a binomial distribution with $\mathbb{P}(\eta(x) = 1) = \mathbb{P}(\eta(x) = 0) = \frac{1}{2}$. Denote $|p - \tilde{p}|$ by $Z$. By Khinchine inequality, $\mathbb{E}[Z] \geq 2e^{\sqrt{\mathbb{E}[Z^2]}} = c/\sqrt{n}$ for some constant $c > 0$. Furthermore, by Paley-Zygmund inequality $\mathbb{P}(Z > \mathbb{E}[Z]/2) \geq \frac{(\mathbb{E}[Z]^2)}{4\mathbb{E}[Z]} \geq c^2$. Hence, with constant probability,

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

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 $\mathcal{H} = \{h_\eta | h_\eta = 1_{\eta(x) \geq \eta}, \eta \in [0, 1]\}$ 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 $\eta$, and hence the maximizer of $\Phi(u, v, p)$ over $\mathcal{H}$ exists due to compactness of the domain of $\eta$. The existence of the maximizer, together with Assumption 1 and TP monotonicity implies by (Narasimhan et al., 2014a, Lemma 11) that $h^*_\text{PU} \in \mathcal{H}$, i.e. the optimal PU classifier is a threshold function on $\eta(x)$ with threshold $\tau^*$, i.e. $h^*_\text{PU} \in \mathcal{H}$.

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

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

$$\Phi(u(h^*_\text{ETU}(x)), v(h^*_\text{ETU}(x)), p) \leq \Phi(u(h^*_\text{PU}), v(h^*_\text{PU}), p) \leq \mathbb{E}_{y|x} \Phi(\tilde{u}(h^*_\text{PU}), \tilde{v}(h^*_\text{PU}), \tilde{p}) + \epsilon/2 \leq \mathbb{E}_{y|x} \Phi(\tilde{u}(h^*_\text{ETU}(x)), \tilde{v}(h^*_\text{ETU}(x)), \tilde{p}) + \epsilon/2,$$

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

$$\Phi(u(h^*_\text{ETU}(x)), v(h^*_\text{ETU}(x)), p) - \Phi(u(h^*_\text{PU}), v(h^*_\text{PU}), p) \leq \epsilon.$$

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

$$\left| \mathbb{E}_{y|x} \Phi(\tilde{u}(h^*_\text{ETU}(x)), \tilde{v}(h^*_\text{ETU}(x)), \tilde{p}) \right| \leq \epsilon,$$

which finishes the proof.

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

The PU-optimal classifier is:

$$h^*_\text{PU} = \arg\max_h \Phi_{\text{ prec}}(u(h), v(h), p) = \arg\max_h \frac{u(h)}{v(h) + \alpha}.$$

**Proposition 2.**

$$h^*_\text{PU}(x) = \begin{cases} 1, & \text{if } x \in X_1, \\ 0, & \text{else} \end{cases}.$$

**Proof.** Note that for the defined $h^*_\text{PU}$ classifier, we have $u(h^*_\text{PU}) = v(h^*_\text{PU}) = \mathbb{P}(X_1)$, and

$$\Phi_{\text{ prec}}(u(h^*_\text{PU}), v(h^*_\text{PU}), p) = \frac{\mathbb{P}(X_1)}{\mathbb{P}(X_1) + \alpha}.$$

Firstly, observe that for any candidate optimal classifier $h'$, it must hold that $h'(x) = 0$ for all $x \in X_3$ (otherwise the metric strictly decreases). Now, suppose there exists a classifier $h' \neq h^*_\text{PU}$ which has strictly higher utility than $h^*_\text{PU}$. Then, it must be that $h'(x) = 1$ for all
We have, $u(h') = \mathbb{P}(X_1) + \mathbb{P}(X_2)(1 - \sqrt{\alpha})$ and $v(h') = \mathbb{P}(X_1) + \mathbb{P}(X_2)$. So:

$$\Phi_{\text{Prec}}(u(h'), v(h'), p) = \frac{\mathbb{P}(X_1) + \mathbb{P}(X_2)(1 - \sqrt{\alpha})}{\mathbb{P}(X_1) + \mathbb{P}(X_2) + \alpha}.$$ 

But for the chosen small value of $\alpha$, 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, $h^*_\text{PU}$ 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 $x = \{x_1, x_2, \ldots, x_n\}$, recall that the ETU-optimal assignments can be computed as:

$$h^*_\text{ETU}(x) = s^* := \arg\max_{s \in \{0, 1\}^n} \mathbb{E}_{y \sim \mathbb{P}(\cdot | x)} \Phi_{\text{Prec}}(s, y).$$

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

$$h^*_\text{ETU}(x_j) = s^*_j = \begin{cases} 1, & \text{if } x \in X_1, \\ 0, & \text{if } x \in X_3. \end{cases}$$

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

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

$$\mathbb{E}_{y \sim \mathbb{P}(\cdot | x)} \Phi_{\text{Prec}}(s^*, y) = \sum_{j \in I_2} s^*_j + \Delta(s^*_j, y_{x_j}) + \sum_{j \in I_2} s^*_j + \sum_{j \in I_2} s^*_j + \alpha n,$$

where $s^*_j$ indicates the optimal assignments corresponding to indices in $I_2$ and $\Delta(s^*_j, y_{x_j})$ is a quantity that depends only on indices in $I_2$, and is given by:

$$\Delta(s^*_j, y_{x_j}) = \sum_{y_{x_j} \in \{0, 1\}^{I_2 \setminus I_2}} \mathbb{P}(y_{I_2}) | (y_{I_2}, s^*_j) | (3)$$

Fixing the optimal predictions for indices corresponding to $I_2$, the value (2) is maximized by maximizing the numerator term $\sum_{j \in I_2} s^*_j$ and minimizing the denominator term $\sum_{j \in I_2 \cup I_3} s^*_j$. This is achieved precisely when the optimal solution satisfies the statement in the proposition. The proof is complete. □

We know from Proposition 2 that $h^*_\text{PU}$ sets the labels corresponding to indices in the set $I_2$ to 0. Now let us examine what happens in the case of ETU, when labels have mild noise (i.e. with some small probability $\sqrt{\alpha}$, the label of an instance from $X_2$ can be 0), at optimality. Consider a candidate optimal solution $s'$ that behaves exactly like $h^*_\text{PU}$, i.e. $s'_j = 0$ for all $j \in I_2$, for some $1 \leq k \leq |I_2|$. Then, $\Delta(s'_{I_2}, y_{I_2}) = 0$, so:

$$\mathbb{E}_{y \sim \mathbb{P}(\cdot | x)} \Phi_{\text{Prec}}(s', y) = \frac{|I_1|}{|I_1| + \alpha n}.$$

Now, consider another candidate solution $s''$ that is equal to $s'$, but has a value of 1 corresponding to a subset of indices $j_1, j_2, \ldots, j_k \in I_2$. The value of this solution can be shown to be:

$$\mathbb{E}_{y \sim \mathbb{P}(\cdot | x)} \Phi_{\text{Prec}}(s'', y) = \frac{|I_1| + k(1 - \epsilon)}{|I_1| + k + \alpha n}.$$

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

$$\epsilon < \frac{\alpha n}{|I_1| + \alpha n},$$

then $s''$ is a strictly better solution than $s'$. In particular, as (5) is monotonic in $k$, the optimal choice is $k = |I_2|$. This immediately leads to the following corollary.

**Corollary 1.**

1. If $|I_2| = 0$, then

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

2. Otherwise, if $\epsilon < \frac{\alpha n}{|I_1| + 1 + \alpha n}$, then

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

In particular, $h^*_\text{ETU}$ assigns label 1 to all instances that are overwhelmingly positive under $\mathbb{P}$, corresponding to indices $I_2$, whereas $h^*_\text{PU}$ assigns label 0. 3. If $|I_1| = 0$, but $|I_2| > 0$ then for any $0 < \epsilon < 1$,

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

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 $\mathbb{P}(X_2)$ is very large.

**C. Proofs for Section 4.1**

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

\[ \hat{u}(h) = \frac{1}{n} \sum_{i=1}^{n} h_i y_i, \quad \hat{v}(h) = \frac{1}{n} \sum_{i=1}^{n} h_i, \quad \hat{\rho} = \frac{1}{n} \sum_{i=1}^{n} y_i, \]

where we used the independence of labels \( y_i, i = 1, \ldots, n \). Similarly, \( E_{y|x} [ (\hat{\rho} - \rho)^2 ] \) is at most \( \frac{1}{n} \), which in total gives:

\[ E_{y|x} \left[ (\hat{\Sigma} - \Sigma)^T \nabla^2 \Phi(\hat{\Sigma} - \Sigma) \right] \leq \frac{A}{n}. \]

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

\[ E_{y|x} \left[ (\hat{\Sigma} - \Sigma)^T \nabla^2 \Phi(\hat{\Sigma} - \Sigma) \right] \geq -\frac{A}{n}. \]

From that we have:

\[ \| E_{y|x} [ \Phi(\hat{\Sigma}) - \Phi(\Sigma) ] \| \leq \frac{A}{2n}, \]

which is exactly what was to be shown.

C.2. Proof of Lemma 3

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

\[ \Phi(\hat{\Sigma}) = \Phi(\hat{\Sigma}) + \nabla \Phi(\hat{\Sigma})^T (\hat{\Sigma} - \Sigma) + \frac{1}{2} (\hat{\Sigma} - \Sigma)^T \nabla^2 \Phi(\hat{\Sigma})(\hat{\Sigma} - \Sigma) \]

for some \( \Sigma \) between \( \hat{\Sigma} \) and \( \Sigma \). Note that \( E_{y|x} [ \hat{\Sigma} ] = \Sigma \), so that:

\[ E_{y|x} \left[ \nabla \Phi(\hat{\Sigma})^T (\hat{\Sigma} - \Sigma) \right] = 0. \]

Furthermore, note that:

\[ (\hat{\Sigma} - \Sigma)^T \nabla^2 \Phi(\hat{\Sigma})(\hat{\Sigma} - \Sigma) \]

\[ = \nabla_{\Sigma \Sigma}^2 (\hat{\Sigma} - \Sigma)^2 + 2 \nabla_{\Sigma \rho}^2 (\hat{\Sigma} - \Sigma)(\hat{\rho} - \rho) + \nabla_{\rho \rho}^2 (\hat{\rho} - \rho)^2 \]

\[ \leq A (\hat{\Sigma} - \Sigma)^2 + 2 |(\hat{\Sigma} - \Sigma)(\hat{\rho} - \rho)| + (\hat{\rho} - \rho)^2 \]

\[ \leq 2A (\hat{\Sigma} - \Sigma)^2 + (\hat{\rho} - \rho)^2, \]

where we used elementary inequality \( ab \leq a^2 + b^2 \), and \( \nabla_{\Sigma \Sigma}^2, \nabla_{\Sigma \rho}^2, \nabla_{\rho \rho}^2 \) denote the second-order derivatives evaluated at some \( \Sigma = (\rho, \rho) \). Hence:

\[ E_{y|x} \left[ (\hat{\Sigma} - \Sigma)^T \nabla^2 \Phi(\hat{\Sigma})(\hat{\Sigma} - \Sigma) \right] \]

\[ \leq 2A \left( E_{y|x} [ (\hat{\Sigma} - \Sigma)^2 ] + E_{y|x} [ (\hat{\rho} - \rho)^2 ] \right). \]

Since \( \hat{\Sigma} \) is the empirical average over \( n \) labels and \( \hat{\Sigma} \) is its expectation (over the labels), \( E_{y|x} [ (\hat{\Sigma} - \Sigma)^2 ] \) is the variance of \( \hat{\Sigma} \), which is at most \( \frac{1}{n^2} \), because \( \hat{\Sigma} \in [0,1] \):

\[ \text{var}(\hat{\Sigma}) = \frac{1}{n^2} \sum_{i=1}^{n} \text{var}(h_i y_i) \leq \frac{1}{n^2} \sum_{i=1}^{n} h_i (1 - h_i) \leq \frac{1}{n}, \]

so that:

\[ \text{tr} \left( \nabla^2 \Phi(\hat{\Sigma}) \Sigma \right) = (\nabla_{\Sigma \Sigma}^2 + 2 \nabla_{\Sigma \rho}^2) s_u + \nabla_{\rho \rho}^2 s_p, \]

Similarly, \( E_{y|x} [ (\hat{\rho} - \rho)^2 ] \) is at most \( \frac{1}{n} \), which in total gives:

\[ E_{y|x} \left[ (\hat{\Sigma} - \Sigma)^T \nabla^2 \Phi(\hat{\Sigma})(\hat{\Sigma} - \Sigma) \right] \leq \frac{A}{n}. \]
where:
\[
 s_p := \sum_{i=1}^{n} \frac{1}{n} \sum_{i=1}^{n} \eta_i (1 - \eta_i),
\]
\[
 s_u := \frac{1}{n^2} \sum_{i=1}^{n} b_i \eta_i (1 - \eta_i),
\]
and \( \nabla_{u u}^2, \nabla_{a p}^2, \nabla_{p p}^2 \) denote the second-order derivative terms evaluated at \((\hat{u}, \hat{p})\). Thus, to finish the proof, we only need to show that the first third-order term is bounded by \( \frac{B}{n} \frac{\nu}{n^{3/2}} \). To this end, note that for any numbers \( a_i, b_{j,k} \), such that \( |b_{j,k}| \leq B, i, j, k = 1, \ldots, 2 \):
\[
 \sum_{i,j,k} b_{i,j} a_i a_j a_k \leq B \sum_{i,j,k} |a_i| |a_j| |a_k| = B (|a_1| + |a_2|)^3.
\]
By Hölder’s inequality,
\[
 \sum_{i=1}^{2} |a_i| \leq \left( \sum_{i=1}^{2} |a_i|^3 \right)^{1/3} 2^{2/3},
\]
so that:
\[
 B (|a_1| + |a_2| + |a_3|)^3 \leq 4B \left( |a_1|^3 + |a_2|^3 + |a_3|^3 \right).
\]
Hence, if we bound:
\[
 \frac{\partial^3 \Phi (z)}{\partial z_\alpha \partial z_\beta \partial z_\gamma} \leq B,
\]
the third-order term \( \frac{2B}{3} \sum_{\alpha, \beta, \gamma=1}^{n} \) is bounded by:
\[
 \frac{2B}{3} \left( |\hat{u} - \bar{u}|^3 + |\hat{p} - \bar{p}|^3 \right).
\]
We now bound \( \mathbb{E}_{y|x} [\hat{u}^3 - \bar{u}^3] \) and \( \mathbb{E}_{y|x} [\hat{p}^3 - \bar{p}^3] \). By Cauchy-Schwarz inequality,
\[
 \mathbb{E}_{y|x} [\hat{p}^3 - \bar{p}^3] \leq \sqrt{\mathbb{E}_{y|x} [\hat{p}^4 - \bar{p}^4]} \sqrt{\mathbb{E}_{y|x} [\hat{p}^2 - \bar{p}^2]}.
\]
Before, we already showed that
\[
 \mathbb{E}_{y|x} [\hat{p}^2 - \bar{p}^2] \leq \frac{1}{4n}.
\]
Denote \( a_i = y_i - \eta_i \), and let \( \mu_k = \mathbb{E}_{y|x} [a_k^4] \). Using \( \mu_1 = 0 \), we have:
\[
 \mathbb{E}_{y|x} [\hat{p}^4 - \bar{p}^4] = \frac{1}{n^4} \sum_{i,j,k,\ell} a_i a_j a_k a_\ell
\]
\[
 = \frac{1}{n^4} \left( n \mu_4 + 3n(n-1) \mu_2^2 \right).
\]
Since \( \mu_2 \leq \frac{1}{4} \) and \( \mu_4 \leq \frac{1}{12} \), \( \mathbb{E}_{y|x} [\hat{p}^4 - \bar{p}^4] \leq \frac{3}{16 n^2} \), and thus:
\[
 \mathbb{E}_{y|x} [\hat{p}^3 - \bar{p}^3] \leq \frac{\sqrt{3}}{8} n^{-3/2} \leq \frac{1}{4} n^{-3/2}.
\]
Using similar bound for \( \mathbb{E}_{y|x} [\hat{u}^3 - \bar{u}^3] \), we conclude that the third-order term is bounded by \( \frac{B}{n} \frac{\nu}{n^{3/2}} \). 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 \( \Phi (h) = \mathbb{E}_{y|x} [\Phi (\hat{u}(h), \hat{v}(h), \hat{p})] \) and \( \Phi_a(h) = \Phi_{\text{appr}}(h) \):
\[
 \Phi (h \text{ETU}) - \Phi_a(h_a) = \Phi (h \text{ETU} - \Phi_a(h_a)) \leq \frac{2B}{2n^{3/2}},
\]
where the bounds shown in the inequalities are from Lemma 3.

**C.4. Derivation of the approximation algorithm for \( F_\beta \)-measure**

Recall that \( F_\beta (u, v, p) = \frac{(1 + \beta^2)u}{\beta^2 p + v} \). The second order derivatives with respect to \( u \) and \( p \) are:
\[
 \frac{\partial^2 F_\beta}{\partial u^2} = 0, \quad \frac{\partial^2 F_\beta}{\partial p^2} = \frac{-\beta^2 (1 + \beta^2)}{(\beta^2 p + v)^3}, \quad \frac{\partial^2 F_\beta}{\partial u \partial p} = \frac{\beta^2 (1 + \beta^2) u}{(\beta^2 p + v)^3}.
\]
To optimize \( \Phi_{\text{appr}}(h) \), we first sort observations according to \( \eta(x_i) \). Then we precompute:
\[
 \hat{p} = \frac{1}{n} \sum_{i=1}^{n} \eta(x_i), \quad \hat{p}_{\text{var}} = \frac{1}{n^2} \sum_{i=1}^{n} \eta(x_i)(1 - \eta(x_i)).
\]
Next, for each \( k = 0, 1, \ldots, n \), we precompute:
\[
 \hat{u}_k = \frac{1}{n} \sum_{i=1}^{k} \eta(x_i), \quad \hat{u}_k = \frac{k}{n}, \quad \hat{u}_{\text{var}} = \frac{1}{n^2} \sum_{i=1}^{k} \eta(x_i)(1 - \eta(x_i)).
\]
We then choose \( k \) for which the ETU approximation:
\[
 \frac{1 + \beta^2 \hat{u}_k}{\beta^2 \hat{p} + \frac{\hat{p}}{n}} = \frac{\beta^2 (1 + \beta^2) \hat{u}_k}{(\beta^2 \hat{p} + \frac{\hat{p}}{n})^3} + \frac{\beta^2 (1 + \beta^2) \hat{u}_k}{(\beta^2 \hat{p} + \frac{\hat{p}}{n})^3} \hat{p}_{\text{var}},
\]
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 \( x = (x_1, \ldots, x_n) \) be the input sample (test set) of size \( n \) generated i.i.d. from \( \mathbb{P} \). Given \( x \) and a function \( \hat{h}: X \rightarrow [0, 1] \), let
\[
 \hat{h} = \arg\max_{h \in \mathbb{H}} \mathbb{E}_{y|x} [\Phi (\hat{u}(h), \hat{v}(h), \hat{p})].
\]

Using similar bound for \( \mathbb{E}_{y|x} [\hat{u}^3 - \hat{u}^3] \), we conclude that the third-order term is bounded by \( \frac{B}{n} \frac{\nu}{n^{3/2}} \). 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.
be the classifier returned by the ETU procedure upon receiving the input sample \(x\). Likewise, let:

\[
h^* = \arg\max_{h \in \hat{\mathcal{H}}} \mathbb{E}_{\eta \sim \eta(x)} \mathbb{E}_{y \sim \eta(h)} \mathbb{E}_{\hat{\mathcal{V}} \sim \hat{\eta}(h)} \left[ \Phi(\hat{u}(h), \hat{v}(h), \hat{\varphi}) \right],
\]

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

\[
\mathbb{E}_x \left[ \Phi_{\text{ETU}}(\hat{h}) - \Phi_{\text{ETU}}(h^*) \right] = \mathbb{E}_x \left[ \Phi_{\text{ETU}}(h^*) \right] - \mathbb{E}_x \left[ \Phi_{\text{ETU}}(\hat{h}) \right] \\
= \mathbb{E}_x \left[ \Phi_{\text{ETU}}(h^*) \right] - \mathbb{E}_x \left[ \Phi_{\text{ETU}}(h^*) \right] + \mathbb{E}_x \left[ \Phi_{\text{ETU}}(h^*) \right] - \mathbb{E}_x \left[ \Phi_{\text{ETU}}(h^*) \right] \\
\leq \mathbb{E}_x \left[ \Phi_{\text{ETU}}(h^*) \right] - \mathbb{E}_x \left[ \Phi_{\text{ETU}}(h^*) \right] \\
\leq 2 \sup_{h \in \hat{\mathcal{H}}} \left| \mathbb{E}_x \left[ \Phi_{\text{ETU}}(h) - \hat{\Phi}_{\text{ETU}}(h) \right] \right|. \tag{7}
\]

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

\[
\Phi_{\text{ETU}}(h) = \mathbb{E}_{y|x} \left[ \Phi(\tilde{u}(h), \tilde{v}(h), \tilde{\varphi}) \right], \\
\tilde{\Phi}_{\text{ETU}}(h) = \mathbb{E}_{y|x} \left[ \Phi(\tilde{u}'(h), \tilde{v}'(h), \tilde{\varphi}) \right]
\]

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

\[
\left| \mathbb{E}_x \left[ \Phi_{\text{ETU}}(h) - \tilde{\Phi}_{\text{ETU}}(h) \right] \right| \\
\leq \mathbb{E} \left[ \left| \Phi(\hat{u}, \hat{v}, \hat{\varphi}) - \Phi(\tilde{u}', \tilde{v}', \tilde{\varphi}) \right| \right] \\
\leq \mathbb{E} \left[ \left| \Phi(\hat{u}, \hat{v}, \hat{\varphi}) - \Phi(u, v, p) \right| \right] \\
+ \mathbb{E} \left[ \left| \Phi(u, v, p) - \Phi(\tilde{u}', \tilde{v}', \tilde{\varphi}) \right| \right],
\]

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 \((x, y)\), and for the sake of clarity we drop the dependence on \(h\) in \(\tilde{u}(h), \tilde{v}(h), \tilde{v}'(h)\). Now, it follows from Lemma 1 that:

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

for some constant \(c\). Moreover, using \(p\)-Lipschitzness of \(\Phi\), we have:

\[
\mathbb{E} \left[ \left| \Phi(u, v, p) - \Phi(\tilde{u}', \tilde{v}', \tilde{\varphi}) \right| \right] \leq U_p \mathbb{E} \left[ |\tilde{\varphi} - \varphi| \right] + V_p \mathbb{E} \left[ |\tilde{v} - v| \right] + P_p \mathbb{E} \left[ |\tilde{p} - p| \right].
\]

Now, the term \(\mathbb{E} \left[ |\tilde{v} - v| \right] \) 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 \(\mathbb{E} \left[ |\tilde{p} - p| \right] \) and \(\mathbb{E} \left[ |\tilde{\varphi} - \varphi| \right] \). Define:

\[
\tilde{p} = \mathbb{E}_{y|x} [\tilde{p}] = \frac{1}{n} \sum_{i=1}^{n} \tilde{\eta}(x_i), \\
\tilde{u}' = \mathbb{E}_{y|x} [\tilde{u}'] = \frac{1}{n} \sum_{i=1}^{n} h(x_i) \tilde{\eta}(x_i), \\
p_{\tilde{\eta}} = \mathbb{E}_x [\tilde{p}] = \mathbb{E} [\tilde{\eta}(x)], \\
u_{\tilde{\eta}} = \mathbb{E}_x [\tilde{v}] = \mathbb{E} [h(x) \tilde{\eta}(x)].
\]

We decompose:

\[
|p - \tilde{p}| \leq |p - p_{\tilde{\eta}}| + |p_{\tilde{\eta}} - \tilde{p}| + |\tilde{p} - p|
\]

As before, we use the fact that \(\mathbb{E}_x [ |p_{\tilde{\eta}} - \tilde{p}| ]\), as well as \(\mathbb{E}_{y|x} [ |\tilde{p} - p| ]\) 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:

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

Using analogous way of reasoning, one gets:

\[
\mathbb{E} \left[ |\tilde{\varphi} - \varphi| \right] \leq |v - v_{\tilde{\eta}}| + \frac{1}{\sqrt{n}}.
\]

Putting it all together, we get:

\[
\mathbb{E}_x \left[ \left| \Phi_{\text{ETU}}(h) - \tilde{\Phi}_{\text{ETU}}(h) \right| \right] \\
\leq c' \sqrt{\frac{\log n}{n}} + U_p |u(h) - u_{\tilde{\eta}}(h)| + P_p |p - p_{\tilde{\eta}}|,
\]

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

\[
\mathbb{E}_x \left[ \left| \Phi_{\text{ETU}}(\hat{h}) - \Phi_{\text{ETU}}(h^*) \right| \right] \leq c' \sqrt{\frac{\log n}{n}} + P_p |p - p_{\tilde{\eta}}| \\
+ \sup_{h \in \hat{\mathcal{H}}} U_p |u(h) - u_{\tilde{\eta}}(h)|,
\]

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 \( w^T x \) (using \( w \) 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: \( w_T, u_T \)

\[
\begin{align*}
& w_0 \leftarrow 0 \\
& u_0 \leftarrow \min(\max(0, 2 \cdot z + 1), 1) \\
& \text{for } t = 1, 2, \ldots, T \text{ do} \\
& \quad w_t \leftarrow w_{t-1} + \frac{1}{n} \sum_{i=1}^{n} (y_i - u_{t-1}(\langle w_{t-1}, x_i \rangle)) \cdot x_i \\
& \quad u_t \leftarrow \text{PAV}\left(\{(w_t, x_i), y_i\}\right)
\end{align*}
\]

end for