A. Proof of Lemma 3

Lemma 3. The stability parameter of a performance measure $\Psi(\cdot)$ can be written as $\delta(\epsilon) \leq L_\psi \cdot \epsilon$ iff its sufficient dual region is bounded in a ball of radius $\Theta(L_\psi)$.

Proof. Let us denote primal variables using the notation $\theta$ proof follows from the fact that any value of $\theta$ for which $\Psi^*(\theta) = -\infty$ can be safely excluded from the sufficient dual region.

For proving the result in one direction suppose $\Psi$ is stable with $\delta(\epsilon) \leq L \epsilon$ for some $L > 0$. Now consider some $\theta \in \mathbb{R}^2$ such that $\|\theta\|_2 \geq L$. Now set $x_C = -C \cdot \theta$. Then we have

$$
\Psi^*(\theta) = \inf_x \{ \langle \theta, x \rangle - \Psi(x) \}
\leq \inf_{C > 0} \{ \langle \theta, x_C \rangle - \Psi(x_C) \}
= \inf_{C > 0} \{ -C \|\theta\|_2^2 - \Psi(x_C) \}
\leq \inf_{C > 0} \{ -C \|\theta\|_2^2 - \Psi(0) + C L \|\theta\|_\infty \}
\leq \inf_{C > 0} \{ -C \|\theta\|_2^2 - \Psi(0) + C L \|\theta\|_2 \}
= \inf_{C > 0} \{ -C \|\theta\|_2 (||\theta||_2 - L) - \Psi(0) \}
\leq \inf_{C > 0} \{ -C \|\theta\|_2 - \Psi(0) \}
= -\infty
$$

Thus, we can conclude that no dual vector with norm greater than $L$ can be a part of the sufficient dual region. This shows that the sufficient dual region is bounded inside a ball of radius $L$. For proving the result in the other direction, suppose the dual sufficient region is indeed bounded in a ball of radius $R$. Consider two points $x_1, x_2$ such that

$$
\theta_1^* = \arg \min_{\theta \in \mathcal{A}_\psi} \{ \langle \theta, x_1 \rangle - \Psi^*(\theta) \}
\theta_2^* = \arg \min_{\theta \in \mathcal{A}_\psi} \{ \langle \theta, x_2 \rangle - \Psi^*(\theta) \}
$$

Now define $f(\theta, x) := \langle \theta, x \rangle - \Psi^*(\theta)$ so that, by the above definition, $f(\theta_1^*, x_1) = \Psi(x_1)$ and $f(\theta_2^*, x_2) = \Psi(x_2)$. Now we have

$$
\Psi(x_1) = f(\theta_1^*, x_1) \leq f(\theta_2^*, x_1)
\leq f(\theta_2^*, x_2) + ||\theta_2^*, x_1 - x_2||
= \Psi(x_2) + ||\theta_2^*, x_1 - x_2||
\leq \Psi(x_2) + R ||x_1 - x_2||_2 ,
$$

where the fourth step follows from the norm bound on $\theta_2^*$. Similarly we have

$$
\Psi(x_2) \leq \Psi(x_1) + R ||x_1 - x_2||_2
$$

This establishes the result.

B. Proof of Theorem 4

Theorem 4. Suppose we are given a stream of random samples $(x_1, y_1), \ldots, (x_T, y_T)$ drawn from a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$. Let $\Psi(\cdot)$ be a concave, Lipschitz link function. Let Algorithm 1 be executed with a dual feasible set $\mathcal{A} \supseteq \mathcal{A}_\psi$, 
Optimizing Non-decomposable Performance Measures: A Tale of Two Classes

\[ \eta_t = 1/\sqrt{t} \text{ and } \eta'_t = 1/\sqrt{t}. \] Then, the average model \( \mathbf{w} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{w}_t \) output by the algorithm satisfies, with probability at least \( 1 - \delta \),

\[ \mathcal{P}_\Psi(\mathbf{w}) \geq \sup_{\mathbf{w}^* \in \mathcal{W}} \mathcal{P}_\Psi(\mathbf{w}^*) - \delta \phi \left( \sqrt{\frac{2B^2}{T}} \log \frac{1}{\delta} \right) - \left( L^2 + 4B^2 \right) \frac{1}{2\sqrt{T}} - \left( L^2 + R^2 \right) \frac{1}{2\sqrt{T}} \sqrt{\log \frac{1}{\delta}}. \]

**Proof.** For this proof we shall assume that \( \Psi \) is \( L \)-Lipschitz so that its sufficient dual region can be bounded by an application of Lemma 3. Notice that the updates for \( (\alpha, \beta) \) can be written as follows:

\[ (\alpha_{t+1}, \beta_{t+1}) \leftarrow \Pi_{A_{\Phi}} \left( (\alpha_t, \beta_t) - \eta_t \nabla_{(\alpha, \beta)} \ell^*_{t}(\alpha_t, \beta_t) \right), \]

where

\[ \ell^*_{t}(\alpha, \beta) \begin{cases} \alpha r^+(\mathbf{w}_t; \mathbf{x}_t, y_t) - \Psi^*(\alpha, \beta) & \text{if } y_t > 0 \\ \beta r^-(\mathbf{w}_t; \mathbf{x}_t, y_t) - \Psi^*(\alpha, \beta) & \text{if } y_t < 0 \end{cases} \]

which can be interpreted as simple gradient descent with \( \ell_t \). Moreover, since \( \Psi^* \) is concave, \( \ell^*_{t}(\alpha, \beta) \) is a \( 2B \)-Lipschitz function. Hence, using a standard GIGA-style analysis (Zinkevich, 2003) on the (descent) updates on \( \alpha_t \) and \( \beta_t \) in Algorithm 1 we have (for \( \eta_t = \frac{1}{\sqrt{t}} \))

\[ \frac{1}{T} \sum_{t=1}^{T} \left[ \alpha_t r^+(\mathbf{w}_t; \mathbf{x}_t, y_t) + \beta_t r^-(\mathbf{w}_t; \mathbf{x}_t, y_t) - \Psi^*(\alpha_t, \beta_t) \right] \]

\[ \leq \inf_{(\alpha, \beta) \in \mathcal{A}} \left\{ \frac{1}{T} \sum_{t=1}^{T} \left[ \alpha r^+(\mathbf{w}_t; \mathbf{x}_t, y_t) + \beta r^-(\mathbf{w}_t; \mathbf{x}_t, y_t) - \Psi^*(\alpha, \beta) \right] \right\} + \left( L^2 + 4B^2 \right) \frac{1}{2\sqrt{T}} \]

\[ = \inf_{(\alpha, \beta) \in \mathcal{A}} \left\{ \alpha \frac{1}{T} \sum_{t=1}^{T} r^+(\mathbf{w}_t; \mathbf{x}_t, y_t) + \beta \frac{1}{T} \sum_{t=1}^{T} r^-(\mathbf{w}_t; \mathbf{x}_t, y_t) - \Psi^*(\alpha, \beta) \right\} + \left( L^2 + 4B^2 \right) \frac{1}{2\sqrt{T}} \]

\[ = \Psi \left( \frac{1}{T} \sum_{t=1}^{T} r^+(\mathbf{w}_t; \mathbf{x}_t, y_t); \frac{1}{T} \sum_{t=1}^{T} r^-(\mathbf{w}_t; \mathbf{x}_t, y_t) \right) + \left( L^2 + 4B^2 \right) \frac{1}{2\sqrt{T}}, \]

where the last step follows from Fenchel conjugacy.

Further, noting that \( \mathbb{E}_{\mathbf{x}_t, y_t} \left[ r^+(\mathbf{w}_t; \mathbf{x}_t, y_t) \right] = P(\mathbf{w}_t) \) and \( \mathbb{E}_{\mathbf{x}_t, y_t} \left[ r^-(\mathbf{w}_t; \mathbf{x}_t, y_t) \right] = N(\mathbf{w}_t) \), we use the standard online-batch conversion bounds (Cesa-Bianchi et al., 2001) to the loss functions \( r^+ \) and \( r^- \) individually to obtain w.h.p.

\[ \frac{1}{T} \sum_{t=1}^{T} r^+(\mathbf{w}_t; \mathbf{x}_t, y_t) \leq \sum_{t=1}^{T} P(\mathbf{w}_t) + \sqrt{\frac{2B^2}{T} \log \frac{1}{\delta}} \]

\[ \frac{1}{T} \sum_{t=1}^{T} r^-(\mathbf{w}_t; \mathbf{x}_t, y_t) \leq \sum_{t=1}^{T} N(\mathbf{w}_t) + \sqrt{\frac{2B^2}{T} \log \frac{1}{\delta}} \]

By monotonicity of \( \Psi \), we get

\[ \frac{1}{T} \sum_{t=1}^{T} \left[ \alpha_t r^+(\mathbf{w}_t; \mathbf{x}_t, y_t) + \beta_t r^-(\mathbf{w}_t; \mathbf{x}_t, y_t) - \Psi^*(\alpha_t, \beta_t) \right] \]

\[ \leq \Psi \left( \frac{1}{T} \sum_{t=1}^{T} P(\mathbf{w}_t); \frac{1}{T} \sum_{t=1}^{T} N(\mathbf{w}_t) \right) + \delta \phi \left( \sqrt{\frac{2B^2}{T} \log \frac{1}{\delta}} \right) + \left( L^2 + 4B^2 \right) \frac{1}{2\sqrt{T}} \]

\[ \leq \Psi \left( \frac{1}{T} \sum_{t=1}^{T} P(\mathbf{w}_t), \frac{1}{T} \sum_{t=1}^{T} N(\mathbf{w}_t) \right) + \delta \phi \left( \sqrt{\frac{2B^2}{T} \log \frac{1}{\delta}} \right) + \left( L^2 + 4B^2 \right) \frac{1}{2\sqrt{T}} \]
\[\frac{1}{T} \sum_{t=1}^{T} \left[\alpha_t r^+(w_t; x_t, y_t) + \beta_t r^-(w_t; x_t, y_t) - \Psi^*(\alpha_t, \beta_t)\right]\]
\[
\geq \frac{1}{T} \sum_{t=1}^{T} \left[\alpha_t P(w^*) + \beta_t N(w^*) - \Psi^*(\alpha_t, \beta_t)\right] - \sqrt{\frac{2L^2_{\Psi}B^2}{T} \log \frac{1}{\delta}} - \left(L^2_{\Psi}L^2_r + R^2_W\right) \frac{1}{2\sqrt{T}}
\]
\[
= \tilde{\alpha} P(w^*) + \beta N(w^*) - \Psi^*(\tilde{\alpha}, \tilde{\beta}) - \sqrt{\frac{2L^2_{\Psi}B^2}{T} \log \frac{1}{\delta}} - \left(L^2_{\Psi}L^2_r + R^2_W\right) \frac{1}{2\sqrt{T}}
\]
\[
\geq \inf_{\alpha, \beta} \left\{\tilde{\alpha} P(w^*) + \beta N(w^*) - \Psi^*(\alpha, \beta)\right\} - \sqrt{\frac{2L^2_{\Psi}B^2}{T} \log \frac{1}{\delta}} - \left(L^2_{\Psi}L^2_r + R^2_W\right) \frac{1}{2\sqrt{T}}
\]
\[
= \Psi(P(w^*), N(w^*)) - \sqrt{\frac{2L^2_{\Psi}B^2}{T} \log \frac{1}{\delta}} - \left(L^2_{\Psi}L^2_r + R^2_W\right) \frac{1}{2\sqrt{T}},
\] where the second step follows from concavity of \(\Psi\) and Jensen's inequality, in the third step \(\tilde{\alpha} = \frac{1}{T} \sum_{t=1}^{T} \alpha_t\) and \(\tilde{\beta} = \frac{1}{T} \sum_{t=1}^{T} \beta_t\), and the last step follows from Fenchel conjugacy.

Combining Eq. (1) and (2) gives us the desired result. 
C. Proof of Theorem 5

Theorem 5. Suppose we have the problem setting in Theorem 4 with the $\Psi_{G\text{-}mean}$ performance measure being optimized for. Consider a modification to Algorithm 1 wherein the reward functions are changed to $r_t^+(\cdot) = r^+(\cdot) + \epsilon(t)$, and $r_t^-(\cdot) = r^-(\cdot) + \epsilon(t)$ for $\epsilon(t) = 1/T^2$. Then, the average model $\overline{w} = \frac{1}{T} \sum_{t=1}^T w_t$ output by the algorithm satisfies, with probability at least $1 - \delta$,

$$
P_{\Psi_{G\text{-}mean}}(\overline{w}) \geq \sup_{w^* \in W} P_{\Psi_{G\text{-}mean}}(w^*) - O \left( \frac{1}{T^{1/4}} \right).$$

Proof. Suppose $\Psi(u + v + \epsilon) \leq \Psi(u, v) + \delta \epsilon(\epsilon)$ as before. Let $r_t^+(\cdot) = r^+(\cdot) + \epsilon(t)$, and $r_t^-(\cdot) = r^-(\cdot) + \epsilon(t)$. Let us make all updates with respect to $r_t^+, r_t^-$. Let $r(\epsilon)$ be the radius of the sufficient dual domain $A$ for a given regularization $\epsilon$. Also let $\bar{\epsilon} = \frac{1}{T} \sum_{t=1}^T \epsilon(t)$. We will assume throughout that $\epsilon(t) = O(1)$. Then we have:

$$\frac{1}{T} \sum_{t=1}^T \left[ \alpha_t r_t^+(w_t; x_t, y_t) + \beta_t r_t^-(w_t; x_t, y_t) - \Psi^*(\alpha_t, \beta_t) \right]$$

$$\leq \inf_{(\alpha, \beta) \in A} \left\{ \frac{1}{T} \sum_{t=1}^T [\alpha_t r_t^+(w_t; x_t, y_t) + \beta_t r_t^-(w_t; x_t, y_t) - \Psi^*(\alpha, \beta)] \right\} + O \left( \frac{r(\bar{\epsilon})}{\sqrt{T}} \right)$$

$$= \inf_{(\alpha, \beta) \in A} \left\{ \alpha \frac{1}{T} \sum_{t=1}^T r^+(w_t; x_t, y_t) + \bar{\epsilon} + \beta \frac{1}{T} \sum_{t=1}^T r^-(w_t; x_t, y_t) + \bar{\epsilon} - \Psi^*(\alpha, \beta) \right\} + O \left( \frac{r(\bar{\epsilon})}{\sqrt{T}} \right)$$

$$= \Psi \left( \frac{1}{T} \sum_{t=1}^T r^+(w_t; x_t, y_t) + \bar{\epsilon}, \frac{1}{T} \sum_{t=1}^T r^-(w_t; x_t, y_t) + \bar{\epsilon} \right) + \delta_{\Psi}(\bar{\epsilon}) + O \left( \frac{r(\bar{\epsilon})}{\sqrt{T}} \right)$$

We can now use online to batch conversion bounds (Cesa-Bianchi et al., 2001), and monotonicity of $\Psi$ to get

$$\frac{1}{T} \sum_{t=1}^T \left[ \alpha_t r_t^+(w_t; x_t, y_t) + \beta_t r_t^-(w_t; x_t, y_t) - \Psi^*(\alpha_t, \beta_t) \right]$$

$$\leq \Psi \left( P(\overline{w}), N(\overline{w}) \right) + \delta_{\Psi} \left( \bar{\Omega} \left( \frac{1}{\sqrt{T}} \right) \right) + \delta_{\Psi}(\bar{\epsilon}) + O \left( \frac{r(\bar{\epsilon})}{\sqrt{T}} \right),$$

For the primal updates, we get, for any $w^* \in W$,

$$\frac{1}{T} \sum_{t=1}^T \left[ \alpha_t r_t^+(w_t; x_t, y_t) + \beta_t r_t^-(w_t; x_t, y_t) - \Psi^*(\alpha_t, \beta_t) \right]$$

$$\geq \frac{1}{T} \sum_{t=1}^T \left[ \alpha_t r_t^+(w^*; x_t, y_t) + \beta_t r_t^-(w^*; x_t, y_t) - \Psi^*(\alpha_t, \beta_t) \right] - \bar{\Omega} \left( \frac{r(\bar{\epsilon})}{\sqrt{T}} \right)$$

$$= \frac{1}{T} \sum_{t=1}^T \left[ \alpha_t r_t^+(w^*; x_t, y_t) + \beta_t r_t^-(w^*; x_t, y_t) - \Psi^*(\alpha_t, \beta_t) \right] + \frac{1}{T} \sum_{t=1}^T \epsilon(t)(\alpha_t + \beta_t) - \bar{\Omega} \left( \frac{r(\bar{\epsilon})}{\sqrt{T}} \right)$$

$$\geq \frac{1}{T} \sum_{t=1}^T \left[ \alpha_t r_t^+(w^*; x_t, y_t) + \beta_t r_t^-(w^*; x_t, y_t) - \Psi^*(\alpha_t, \beta_t) \right] - \bar{\Omega} \left( \frac{r(\bar{\epsilon})}{\sqrt{T}} \right),$$

since $\epsilon(t), \alpha_t, \beta_t \geq 0$. Again using an online-batch conversion argument (Cesa-Bianchi et al., 2001) we get w.h.p,

$$\frac{1}{T} \sum_{t=1}^T \left[ \alpha_t r_t^+(w_t; x_t, y_t) + \beta_t r_t^-(w_t; x_t, y_t) - \Psi^*(\alpha_t, \beta_t) \right] \geq \Psi(\overline{P}(w^*), N(\overline{w}^*)) - \bar{\Omega} \left( \frac{r(\bar{\epsilon})}{\sqrt{T}} \right).$$
Combining Eq. (4) and (5) gives us
\[ \Psi(P(w), N(w)) \geq \Psi(P(w^*), N(w^*)) - \tilde{O} \left( \frac{r(\bar{\epsilon})}{\sqrt{T}} \right) - \delta \left( \frac{1}{\sqrt{T}} \right) - \delta(\bar{\epsilon}) \]

For G-mean, \( \delta(\bar{\epsilon}) = \sqrt{\bar{\epsilon}} \), and by an application of Lemma 3, we have \( r(\epsilon) = O(\frac{1}{\sqrt{\epsilon}}) \). Thus we have
\[ \Psi(P(w), N(w)) \geq \Psi(P(w^*), N(w^*)) - \tilde{O} \left( \frac{1}{\sqrt{T}} \right) - \sqrt{\bar{\epsilon}} \]

For \( \bar{\epsilon} = O(\frac{1}{4\sqrt{T}}) \), we get
\[ \Psi(P(w), N(w)) \geq \Psi(P(w^*), N(w^*)) - \tilde{O} \left( \frac{1}{4\sqrt{T}} \right) \]

This can be achieved with \( \epsilon(t) = \frac{1}{\sqrt{T}} \).

**D. Proof of Theorem 8**

**Theorem 8.** Let Algorithm 2 be executed with a performance measure \( P(a, b) \) and reward functions that offer values in the range \([0, m]\). Let \( P^* := \sup_{w \in W} P(a, b)(w) \). Also let \( \Delta_t = P^* - P(a, b)(w_t) \) be the excess error for the model \( w_t \) generated at time \( t \). Then there exists a value \( \eta(m) < 1 \) such that for \( \Delta_t \leq \Delta_0 \cdot \eta(m)^t \).

**Proof.** In order to be generic in its treatment, the proof will require the following regularity conditions on the performance measure

1. \( b_0 \neq 0 \)
2. \( \alpha - P(w) \cdot \gamma \geq 0 \) for all \( w \in W \)
3. \( \beta - P(w) \cdot \delta \geq 0 \) for all \( w \in W \)
4. \( -1 < f \leq \gamma \cdot P(w) + \delta \cdot N(w) \leq g \) for all \( w \in W \)

Define \( e_t := V(w_{t+1}, v_t) - v_t \). Then we can state the following lemmata which together yield the convergence bound proof.

**Lemma 10.** \( \frac{e_t}{1+f} \geq P^* - v_t \)

**Proof.** Assume that for some \( w^* \), \( P(w^*) = v_t + e_t + e' \) where \( e' > 0 \). Then we have
\[ V(w^*, v_t) = \left( \frac{e_t}{1+f} + e' \right) (1 + \gamma \cdot P(w^*) + \delta \cdot N(w^*)) - e_t \]
\[ \geq \left( \frac{e_t}{1+f} + e' \right) (1 + f) - e_t \]
\[ = e'(1 + f) > 0, \]

which contradicts the fact that no classifier can achieve a valuation greater than \( v_t + e_t \) at level \( v_t \), thus proving the desired result.

**Lemma 11.** For any \( w \) that achieves \( V(w, v) = v + e \) such that \( e \geq 0 \), we have
\[ P(w) \geq v + \frac{e}{g+1} \]
Proof. Let \( v' = v + \frac{c}{g+1} \). We will show that \( V(w, v') \geq v' \) which will establish the result by pseudo-linearity. We have

\[
V(w, v') - v' = c + (\alpha - v' \gamma) \cdot P(w) + (\beta - v' \delta) \cdot N(w) - v'
\]

\[
= c + (\alpha - v \gamma) \cdot P(w) + (\beta - v \delta) \cdot N(w) - v' - \frac{c}{g+1} (\gamma \cdot P(w) + \delta \cdot N(w))
\]

\[
= v + e - v' - \frac{c}{g+1} (\gamma \cdot P(w) + \delta \cdot N(w))
\]

\[
\geq v + e - v' - \frac{g e}{g+1} = 0,
\]

where we have used the bounds on \( \gamma \cdot P(w) + \delta \cdot N(w) \) and the fact that \( 1 + g > 0 \).

Given the above results we can establish the convergence bound. More specifically, we can show the following: let \( \Delta_t = P^* - P(w_t) \). Then we have

\[
\Delta_{t+1} \leq \frac{g - f}{g + 1} \cdot \Delta_t
\]

To see this, consider the following

\[
\Delta_{t+1} = P^* - P(w_{t+1}) \leq P^* - \left( v_t + \frac{e_t}{g+1} \right) \leq P^* - \left( v_t + \frac{1 + f}{g+1} \cdot (P^* - v_t) \right)
\]

\[
= P^* - \left( P(w_t) + \frac{1 + f}{g+1} \cdot (P^* - P(w_t)) \right) = \Delta_t - \frac{1 + f}{g+1} \cdot \Delta_t = \frac{g - f}{g + 1} \cdot \Delta_t,
\]

which proves the result. Notice that Table 2 gives the rates of convergence for the different performance measures by calculating bounds on the value of \( \frac{g - f}{g + 1} \) for those performance measures.

E. An analysis of the AMP Algorithm under Inexact Maximizations

For this and the next section, we will, for the sake of simplicity, we will focus only on the F-measure for \( \beta = 1 \) and \( p = 1/2 \) so that \( \theta = 1 \). For this setting, the F-measure looks like the following: \( F(P, N) = \frac{2P}{2P - N} \), and the valuation function looks like \( V(w, v) = (1 - v/2) \cdot P(w) + v/2 \cdot N(w) \). We shall denote the performance measure as \( F(w) \), and its optimal value as \( F^* \). We will assume that the reward functions give bounded rewards in the range \([0, m]\).

So far we assumed that Step 4 in the Algorithm AMP gave us \( w_{t+1} \) such that

\[
V(w_{t+1}, v_t) = \max_{w \in \mathcal{W}} V(w, v_t)
\]

Now we will only assume that \( w_{t+1} \) satisfies

\[
V(w_{t+1}, v_t) = \max_{w \in \mathcal{W}} V(w, v_t) - \epsilon_t
\]

We also assume that the level \( v_t \) is only approximated in Step 5 of AMP, i.e. using Lemma 7 we have

\[
v_t = F(w_t) + \delta_t
\]

where \( \delta_t \) is a signed real number.

Given these approximations, we can prove the following results

**Lemma 12.** The following hold for the setting described above

1. If \( \delta_t \leq 0 \) then \( e_t \geq 0 \)
2. If \( \delta_t > 0 \) then \( e_t \geq -\delta_t (1 + \frac{m}{2}) \)
3. If \( F^* < v_t \) (which can happen only if \( \delta_t > 0 \), then \( e_t < 0 \)
4. If \( e_t < 0 \) then \( F^* < v_t \)
5. We have
   
   
   (a) If $e_t \geq 0$, then $e_t \geq \left( \frac{2-m}{2} \right) (F^* - v_t)$. 
   
   (b) If $e_t < 0$, then $e_t \geq \left( \frac{2+m}{2} \right) (F^* - v_t)$. 
   
6. If $V(w, v) = v + e$, then
   
   (a) If $e \geq 0$ then $F(w) \geq v + \frac{2e}{2+m}$ 
   
   (b) If $e < 0$ then $F(w) \geq v + \frac{2e}{2-m}$ 

Proof. We give the proof in parts

1. If $\delta_t \leq 0$ then this means that there exists a $w$ such that $F(w) \geq v_t$. The result then follows from pseudo linearity.

2. $v_t = F(w_t) + \delta_t$ gives us, by pseudo linearity of F-measure,

   $$(1 - v_t/2) \cdot P(w_t) + v_t/2 \cdot N(w_t) = v_t - \delta_t \left( 1 + \frac{P(w_t) - N(w_t)}{2} \right) \geq v_t - \delta_t \left( 1 + \frac{m}{2} \right).$$

The bound on $e_t$ now follows from its definition.

3. Suppose $e_t \geq 0$ then by pseudo linearity of F-measure, we have, for some $w$, $V(w, v_t) \geq v_t$ which contradicts the assumption.

4. Suppose there exists $w^*$ with $F(w^*) = v_t + e'$ with $e' \geq 0$ then we have

   $$(1 - v_t/2) \cdot P(w^*) + v_t/2 \cdot N(w^*) = v_t + e' \left( 1 + \frac{P(w^*) - N(w^*)}{2} \right) \geq 0,$$

   which contradicts the fact that $e_t < 0$.

5. Part (a) is simply Lemma 10. For part (b), we will prove that $F^* \leq v_t + \frac{2e_t}{2+m}$. Since $\frac{2}{2+m} > 0$, the result will follow.

   Assume the contrapositive that some $w^*$ achieves $F(w^*) = v_t + \frac{2e_t}{2+m} + e'$ for some $e' > 0$. Using the pseudo linearity of F-measure (and using the shorthand $v' = v_t + \frac{2e_t}{2+m} + e'$), this can be expressed as

   $$(1 - v'/2) \cdot P(w^*) + v'/2 \cdot N(w^*) = v'$$

   where for some $e' > 0$. Then we have

   $$(1 - v_t/2) \cdot P(w^*) + v_t/2 \cdot N(w^*) - v_t - e_t = v' - v_t - e_t + \frac{1}{2} \left( \frac{2e_t}{2+m} + e' \right) (P(w^*) - N(w^*))$$

   $$= \frac{2e_t}{2+m} + e' - e_t + \frac{1}{2} \left( \frac{2e_t}{2+m} + e' \right) (P(w^*) - N(w^*))$$

   $$\geq \frac{2e_t}{2+m} + e' - e_t + \frac{m}{2} \left( \frac{2e_t}{2+m} + e' \right)$$

   $$= e' \left( 1 + \frac{m}{2} \right) + e_t \left( \frac{2}{2+m} - 1 + \frac{m}{2+m} \right)$$

   $$= e' \left( 1 + \frac{m}{2} \right) > 0,$$

   where we have assumed that $e'$ is chosen small enough so that $\frac{2e_t}{2+m} + e' < 0$ still and used the fact that $P(w^*) - N(w^*) \leq m$.

6. Part (a) is simply Lemma 11. To prove part (b), we let $v' = v + \frac{2e}{2-m}$, then we have

   $$(1 - v'/2) \cdot P(w) + \frac{v'}{2} \cdot N(w) - v' = (1 - v/2) \cdot P(w) + \frac{v}{2} \cdot N(w) - v + \frac{e}{2-m} (N(w) - P(w))$$
Optimizing Non-decomposable Performance Measures: A Tale of Two Classes

\[ \begin{align*}
&\geq (1 - \frac{v}{2}) \cdot P(w) + \frac{v}{2} \cdot N(w) - v' + \frac{me}{2 - m} \\
&= v + e - \left( v + \frac{2e}{2 - m} \right) + \frac{me}{2 - m} \\
&= e \left( 1 - \frac{2}{2 - m} + \frac{m}{2 - m} \right) \\
&= 0,
\end{align*} \]

where the second inequality follows since \( N(w) - P(w) \leq m \) and \( e < 0 \) by using the bounds on the reward functions. This proves the result.

\[ \square \]

E.1. Convergence analysis

We have the following cases with us

1. Case 1 (\( \delta_t \leq 0 \)): In this case we are setting \( v_t \) to a value less than the F-measure of the current classifier. This should hurt performance - we know that \( v_t = F(w_t) + \delta_t \) which gives us, on applying part (a) of the previous lemma using \( F^* - v_t = \Delta_t - \delta_t \), the following

\[ e_t \geq \frac{2 - m}{2} (\Delta_t - \delta_t). \]

Note that we are guaranteed that \( e_t \geq 0 \) in this case. Now since the maximization in step 4 is also carried out approximately, we have \( V(w_{t+1}, v_t) = v_t + e_t - \epsilon_t \). Now we have two sub cases

(a) Case 1.1 (\( \epsilon_t \leq e_t \)): In this case we can apply part 6(a) of the previous lemma to get the following result

\[ \Delta_{t+1} \leq \frac{2m}{2 + m} \Delta_t - \frac{2m}{2 + m} \delta_t + \frac{2\epsilon_t}{2m} \]

(b) Case 1.2 (\( \epsilon_t > e_t \)): In this case we are actually making negative progress in the maximization step (since we have \( V(w_{t+1}, v_t) \leq v_t \) and we can only invoke Lemma 5.6(b) to get

\[ \Delta_{t+1} \leq \frac{2\epsilon_t}{2 - m} \]

Note that the above result should not be interpreted as a one shot step to a very good classifier. The above result holds along with the condition that \( \epsilon_t > e_t \). Thus the performance of the classifier is lower bounded by \( e_t \) which depends on how far the current classifier is from the best.

2. Case 2 (\( \delta_t > 0 \)): In this case we are setting \( v_t \) to the value higher than the F-measure of the current classifier. This can mislead the classifier and results in the following two sub-cases

(a) Case 2.1 (\( F^* \geq v_t \)): In this case we are still setting \( v_t \) to a legitimate value, i.e. one that is a valid F-measure for some classifier in the hypothesis class. This can only benefit the next optimization stage (in fact if we set \( v_t = F^* \), then we would obtain the best classifier in this very iteration!). In this case \( e_t \geq 0 \) and we can use the analyses of Cases 1.1 and 1.2.

(b) Case 2.2 (\( F^* < v_t \)): In this case we are setting \( v_t \) to an illegal value, one that is an unachievable value of F-measure. Consequently, using part 3 of the previous lemma, \( e_t < 0 \) and using part(b) of the previous lemma we get

\[ e_t \geq \frac{2 + m}{2} (\Delta_t - \delta_t), \]

which, upon applying part 6(b) of the previous lemma (since \( e_t - \epsilon_t \leq e_t < 0 \)) will give us

\[ \Delta_{t+1} \leq \frac{2m}{2 - m} (\delta_t - \Delta_t) + \frac{2\epsilon_t}{2 - m} \leq \frac{2m}{2 - m} \delta_t + \frac{2\epsilon_t}{2 - m}. \]
We can combine the cases together as follows

\[ \Delta_{t+1} \leq \max \left\{ \frac{2m}{2+m} \Delta_t - \frac{2m}{2+m} \delta_t + \frac{2\epsilon_t}{2+m} \right\}, \frac{2m}{2+m} \Delta_t + \frac{2\epsilon_t}{2-m} \right\}, \frac{2m}{2+m} \Delta_t + \frac{2\epsilon_t}{2-m} \right\} \}

\[ \leq \max \left\{ \frac{2m}{2+m} \Delta_t + \frac{2m}{2+m} \left| \delta_t \right| + \frac{2\epsilon_t}{2+m}, \frac{2m}{2-m} \left| \delta_t \right| + \frac{2\epsilon_t}{2-m} \right\} \}

If we let \( \eta = \frac{2m}{2+m}, \eta' = \frac{2m}{2-m} \), and \( \xi_t = \left| \delta_t \right| + \epsilon_t/m \), then this gives us

\[ \Delta_{t+1} \leq \eta \Delta_t + \eta' \xi_t, \]

which gives us

\[ \Delta_T \leq \eta^T \Delta_0 + \frac{\eta'}{\eta} \sum_{i=0}^{T-1} \eta^{T-i} \xi_i \]

This concludes our analysis.

F. Proof of Theorem 9

**Theorem 9.** Let Algorithm 3 be executed with a performance measure \( P_{(a,b)} \) and reward functions with range \((0, m)\). Let \( \eta = \eta(m) \) be the rate of convergence guaranteed for \( P_{(a,b)} \) by the AMP algorithm. Set the epoch lengths to \( s_e, s'_e = \tilde{O} \left( \frac{1}{m^2} \right) \). Then after \( e = \log \frac{1}{2} \left( \frac{1}{2} \log^2 \frac{1}{\epsilon} \right) \) epochs, we can ensure with probability at least \( 1 - \delta \) that \( P^* - P_{(a,b)}(w_e) \leq \epsilon \).

Moreover the number of samples consumed till this point is at most \( \tilde{O} \left( \frac{1}{\epsilon^2} \right) \).

**Proof.** Using Hoeffding’s inequality, standard regret and online-to-batch guarantees (Cesa-Bianchi et al. 2001; Zinkevich 2003), we can ensure that, if the stream lengths for the Model optimization stage and Challenge level estimation stage procedures are \( s_e \) and \( s'_e \) respectively, then for some fixed \( c > 0 \) that is independent of the stream length, we have

\[ |\delta_t| \leq c \cdot \sqrt{\frac{\log \frac{1}{2}}{s'_e}}, |\epsilon_t| \leq c \sqrt{\frac{\log \frac{1}{2}}{s_e}} \]

Let \( T = \log \frac{1}{2} \left( \frac{1}{2} \log^2 \frac{1}{\epsilon} \right) \) and \( s_e = (\frac{2c}{m})^2 \left( \frac{1}{\eta} \right)^{2c} \log \frac{T}{\delta} \) and \( s'_e = 4c^2 \left( \frac{1}{\eta} \right)^{2c} \log \frac{T}{\delta} \) - this gives us, for each \( \epsilon \), with probability at least \( 1 - \delta/T \),

\[ \xi_e \leq \eta^c \]

Thus, using a union bound, with probability at least \( 1 - \delta \), we have, by the discussion in the previous section,

\[ \Delta_T \leq \eta^T \Delta_0 + \frac{\eta'}{\eta} \sum_{i=0}^{T-1} \eta^{T-i} \xi_i \leq \eta^T \Delta_0 + \frac{\eta'}{\eta} T \eta^T \]

\[ \leq \epsilon \Delta_0 \log^{-2} \left( \frac{1}{\epsilon} \right) + \frac{\eta'}{\eta} \log \left( \frac{1}{\epsilon} \log^2 \frac{1}{\epsilon} \right) \epsilon \log^{-2} \left( \frac{1}{\epsilon} \right) \]

\[ \leq \epsilon \left( \Delta_0 + \frac{\eta'}{\eta \log \frac{1}{\epsilon}} \right), \]

where the last step follows from the fact that for any \( \epsilon < 1/\epsilon^2 \), we have

\[ \log \left( \frac{1}{\epsilon} \log^2 \frac{1}{\epsilon} \right) \leq \log^2 \left( \frac{1}{\epsilon} \right) \]
Let \( d = \left( \Delta_0 + \frac{\eta'}{\eta \log \frac{1}{\eta}} \right) \) so that we can later set \( \epsilon' = \epsilon / d \), and \( s = 4 \epsilon^2 \left( 1 + \frac{1}{m^2} \right) \) so that \( s_e + s'_e = s \left( \frac{1}{\eta} \right)^{2\epsilon} \log \frac{T}{\delta} \). The total number of samples required can then be calculated as

\[
\sum_{e=1}^{T} s_e + s'_e = s \log \frac{T}{\delta} \sum_{e=1}^{T} \left( \frac{1}{\eta} \right)^{2\epsilon} \leq s \log \frac{T}{\delta} \frac{1}{1 - \eta^2} \frac{1}{\epsilon^2} \log^4 \frac{1}{\epsilon} \leq s \log \frac{T}{\delta} \frac{1}{1 - \eta^2} \frac{1}{
\epsilon^2} \log^4 \frac{1}{\epsilon}
\]

This gives the number of samples required as

\[
O \left( \frac{1}{\epsilon^2} \log^4 \frac{1}{\epsilon} \left( \log \log \frac{1}{\epsilon} + \log \frac{1}{\delta} \right) \right),
\]

to get an \( \epsilon \)-accurate solution with confidence \( 1 - \delta \).