Active Approximately Metric-Fair Learning (Supplementary Material)

Yiting Cao$^1$  
Chao Lan$^1$

$^1$School of Computer Science, University of Oklahoma, Norman, Oklahoma, USA

**Lemma 0.1 (Lemma 3.5).** Fix any $t, \beta > 0$. Let $F : X \times X \rightarrow \mathbb{R}$ be a hypothesis class induced from $H$ such that $\forall f \in F$, $f(x, x') = \tau_{\beta}^t(\langle h(x) - h(x') \rangle)$ where $\tau_{\beta}^t(z)$ is a piecewise model outputting 1 if $z > \beta + \frac{1}{t}$, outputting 0 if $z \leq \beta$ and $t(z - \beta)$ otherwise. Then $\mathcal{R}_m(F) \leq 8t \cdot \mathcal{R}_m(H)$.

**Proof.** Let $G : X \times X \rightarrow \mathbb{R}$ be the set of functions induced from $h$ and defined as $\forall g \in G, g(a, b) = h(a) - h(b)$. Let $\text{abs}$ be the absolute function. Then $f(a, b) = \tau_{\beta}^t \circ \text{abs} \circ g(a, b)$ and we can write, accordingly,

$$F = \tau_{\beta}^t \circ \text{abs} \circ G. \quad (1)$$

We first show $\mathcal{R}_m(F) \leq \mathcal{R}_m(G)$. This is true because

$$\mathcal{R}_m(F) = \mathcal{R}_m(\tau_{\beta}^t \circ \text{abs} \circ G) \leq 2t \cdot \mathcal{R}_m(\text{abs} \circ G) \leq 4t \cdot \mathcal{R}_m(G), \quad (2)$$

where both inequalities are by the property of Rademacher complexity for composite function with one component being Lipschitz continuous e.g., [Bartlett and Mendelson 2002 Theorem 12] and the facts that $\tau_{\beta}^t$ and $\text{abs}$ are both Lipschitz with constants $t$ and 1 respectively.

We then show $\mathcal{R}_m(G) \leq 2 \cdot \mathcal{R}_m(H)$. This is true because

$$\mathcal{R}_m(G) = E_{(a_i, b_i)} E_\sigma \sup_{g \in G} \frac{1}{m} \sum_{i=1}^m \sigma_i g(a_i, b_i)$$

$$= E_{(a_i, b_i)} E_\sigma \sup_{g \in G} \frac{1}{m} \sum_{i=1}^m \sigma_i [h(a_i) - h(b_i)]$$

$$\leq E_{(a_i, b_i)} E_\sigma \sup_{g \in G} \frac{1}{m} \sum_{i=1}^m \sigma_i [h(a_i) + \epsilon_i h(b_i)]$$

$$= E_{(a_i, b_i)} E_\sigma \sup_{g \in G} \frac{1}{m} \sum_{i=1}^m \sigma_i h(x_i)$$

$$= 2 \cdot \mathcal{R}_m(H). \quad (3)$$

where the third equality is based on the fact that $\sigma_i$ is uniform in $\{-1, 1\}$ so the expectation with respect to $\sigma_i$ is the same as the expectation with respect to $-\sigma_i$.

Combining (2) and (3) proves the lemma. \qed

**Theorem 0.2 (Theorem 3.6).** Fix any $\alpha, \beta, t > 0$. Suppose $\mathcal{R}_m(H) \in O(1/\sqrt{m})$. Any model $h \in H$ returned by the AMF learner satisfies $\Delta_{\alpha, \beta+1/t}(h) \leq \epsilon$ with probability at least $1 - \delta$ if $m \geq \frac{1}{\epsilon^2} \left(16tc + \sqrt{\frac{\epsilon}{2 \log \frac{1}{\delta}}} \right)$, where $m$ is the number of $(x, x') \in S$ satisfying $d(x, x') \leq \alpha$ and $c$ is a constant inherited from $O(1/\sqrt{m})$. 

*Accepted for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).*
Proof. To facilitate discussion, define two functions

$$
\tau_\beta(z) = \begin{cases} 
1, & \text{if } z > \beta \\
0, & \text{if } z \leq \beta
\end{cases}
$$

and

$$
\tau_\beta^1(z) = \begin{cases} 
1, & \text{if } z > \beta + \frac{1}{\tau} \\
t(z - \beta), & \text{if } \beta < z \leq \beta + \frac{1}{\tau} \\
0, & \text{if } z \leq \beta
\end{cases}
$$

By definition, we have

$$
\tau_{\beta + \frac{1}{\tau}}(z) \leq \tau_\beta^1(z) \leq \tau_\beta(z).
$$

Recall $S = \{(x_i, x_j)\}_{i,j=1,...,n}$. Let $S_\alpha$ be a subset of $S$ defined as

$$
S_\alpha = \{(a, b) \in S \mid d(a, b) \leq \alpha\}.
$$

Suppose the size of $S_\alpha$ is $m$. Then,

$$
\Delta_{\alpha,\beta}(h; S) = \frac{1}{n^2} \sum_{i,j=1}^n \mathbb{I}\{|h(x_i) - h(x_j)| > \beta, d(x_i, x_j) \leq \alpha\}
$$

$$
= \frac{1}{m^2} \frac{1}{m} \sum_{(a, b) \in S_\alpha} \mathbb{I}\{|h(a) - h(b)| > \beta\}
$$

$$
= \frac{1}{m^2} \cdot \frac{1}{m} \sum_{(a, b) \in S_\alpha} \tau_\beta(|h(a) - h(b)|).
$$

Recall $F : X \times X \to \mathbb{R}$ is the set of functions induced from $\tau_\beta^1$ and defined as $\forall f \in F$, $f(a, b) = \tau_\beta^1(|h(a) - h(b)|)$. We have that, with probability at least $1 - \delta$,

$$
\frac{1}{m} \sum_{(a, b) \in S_\alpha} \tau_\beta(|h(a) - h(b)|) \geq \frac{1}{m} \sum_{(a, b) \in S_\alpha} \tau_\beta^1(|h(a) - h(b)|)
$$

$$
\geq \mathbb{E}[\tau_\beta^1(|h(a) - h(b)|) \mid d(a, b) \leq \alpha] - 2mR_m(F) - \sqrt{\frac{\log \frac{1}{\delta}}{2m}}
$$

$$
\geq \mathbb{E}[\tau_{\beta + \frac{1}{\tau}}(|h(a) - h(b)|) \mid d(a, b) \leq \alpha] - 16tR_m(H) - \sqrt{\frac{\log \frac{1}{\delta}}{2m}}
$$

$$
\geq \mathbb{E}[\tau_{\beta + \frac{1}{\tau}}(|h(a) - h(b)|) \mid d(a, b) \leq \alpha] - \frac{1}{\sqrt{m}} \left(16tc + \sqrt{\frac{1}{2} \log \frac{1}{\delta}}\right).
$$

where for some constant $c$. In (9), the first inequality is by (6); the second one is by standard generalization bound with Rademacher complexity e.g. [Mohri et al., 2018] Theorem 3.3] conditioned on $d(a, b) \leq \alpha$; the third one is by (6) and Lemma 3.5; and the last one holds since $R_m \in O(1/\sqrt{m})$. Note the expectation of $(a, b) \in S_\alpha$ in $R_m \in O(1/\sqrt{m})$ is also conditioned on $d(a, b) \leq \alpha$, and we always assume $R_m \in O(1/\sqrt{m})$ w.r.t. any proper data distribution.

Combining (8) and (9), we see $\Delta_{\alpha,\beta}(h; S) = 0$ implies

$$
\mathbb{E}[\tau_{\beta + \frac{1}{\tau}}(|h(a) - h(b)|) \mid d(a, b) \leq \alpha] \leq \frac{1}{m} \left(16tc + \sqrt{\frac{1}{2} \log \frac{1}{\delta}}\right).
$$

\footnote{Here we follow [Yona and Rothblum, 2018] and treat $S_\alpha$ as an i.i.d. sample. If it is not, we can either add an additional constraint that no two pairs in $S_\alpha$ share the same instance so it can be viewed as an i.i.d. sample, or apply a generalization error bound on non-i.i.d. sample e.g. [Mohri and Rostamizadeh, 2008]. In either case, the order of our result remains the same.}
Further, we can show
\[ \Delta_{\alpha,\beta+\frac{1}{t}}(h) \leq \mathbb{E}[\tau_{\beta+\frac{1}{t}}(|h(a) - h(b)|) \mid d(a, b) \leq \alpha], \] (11)
because
\[
\Delta_{\alpha,\beta+\frac{1}{t}}(h) = \int_{(a,b) \in X \times X} \mathbb{I}\{|h(a) - h(b)| > \beta + 1/t\} \cdot \mathbb{I}\{d(a, b) \leq \alpha\} \cdot p(a, b) \\
\leq \int_{(a,b) \in X \times X} \mathbb{I}\{|h(a) - h(b)| > \beta + 1/t\} \cdot p(a, b) \\
\leq \int_{(a,b) \in X \times X} \mathbb{I}\{|h(a) - h(b)| > \beta + 1/t\} \cdot p(a, b \mid d(a, b) \leq \alpha) \\
= \mathbb{E}[\tau_{\beta+\frac{1}{t}}(|h(a) - h(b)|) \mid d(a, b) \leq \alpha].
\] (12)

Combining (10) and (11), and upper bounding the RHS of (10) by \( \varepsilon \) implies that \( \Delta_{\alpha,\beta+\frac{1}{t}}(h) \leq \varepsilon \) whenever
\[
m \geq \frac{1}{\varepsilon^2} \left( 16tc + \sqrt{\frac{1}{2} \log \frac{1}{\delta}} \right).
\] (13)

The theorem is proved. \( \Box \)

**Theorem 0.3** (Theorem 4.2). Fix any \( \alpha, \beta > 0 \). Suppose \( R_m(H) \in O(1/\sqrt{m}) \) and the counter \((\alpha, \beta)\) AMF coefficient w.r.t. \( H \) is bounded. Then, with probability at least \( 1 - \delta \), any \( h \in H \) returned by Algorithm 1 satisfies \( \Delta_{\alpha,\beta}(h) \leq \varepsilon \) after \( O(\log \frac{1}{\delta}) \) labeling.

**Proof.** Suppose we have performed \( q \) rounds of labeling. Let \( L_q \) be the updated training set and \( S_q \) be the associated set of instance pairs in Definition 3.4. Define
\[ V_q = \{ h \in H; \Delta_{\alpha,\beta}(h; S_q) = 0 \}. \] (14)
Consider labeling \( m \) instances in round \( q + 1 \). First, note that all labeled instances fall in \( C_{\alpha,\beta}(V_q) \) and thus will add to \( S_q \) at least \( m \) pairs of \((x, x')\) satisfying \( d(x, x') \leq \alpha \). Then, by Theorem 0.2 and setting \( t = 1/\beta \), if
\[ m \geq \frac{1}{4xt} \left( 32c/\beta + \sqrt{\frac{1}{2} \log \frac{1}{\delta}} \right), \]
with probability at least \( 1 - \delta' \), any \( h \in V_{q+1} \) satisfies
\[ \Delta_{\alpha,\beta}(h) \leq 1/(2\xi). \] (15)
Let \& be logic ‘AND’ and define event
\[ I^\beta_{\alpha}(x, x'; h) := d(x, x') \leq \alpha \& |h(x) - h(x')| > \beta. \] (16)
Then, with probability at least \( 1 - \delta' \), any \( h \in V_{q+1} \) satisfies
\[
\Pr\{I^\beta_{\alpha}(x, x'; h)\} = \Pr\{I^\beta_{\alpha}(x, x'; h) \& (x, x') \in C_{\alpha,\beta}(V_q)\} + \Pr\{I^\beta_{\alpha}(x, x'; h) \& (x, x') \notin C_{\alpha,\beta}(V_q)\} \\
= \Pr\{I^\beta_{\alpha}(x, x'; h) \& (x, x') \in C_{\alpha,\beta}(V_q)\} \\
= \Pr\{I^\beta_{\alpha}(x, x'; h) \mid (x, x') \in C_{\alpha,\beta}(V_q)\} \cdot \Pr\{(x, x') \in C_{\alpha,\beta}(V_q)\} \\
\leq \frac{\Pr\{(x, x') \in C_{\alpha,\beta}(V_q)\}}{2\xi},
\] (17)
where the second equality is by the fact that \( \Pr\{I^\beta_{\alpha}(x, x'; h) \& (x, x') \notin C_{\alpha,\beta}(V_q)\} \leq \Pr\{I^\beta_{\alpha}(x, x'; h) \& (x, x') \notin C_{\alpha,\beta}(V_{q+1})\} = 0 \), and the inequality is by (15) conditioned on an additional fact that all labeled instances fall in \( C_{\alpha,\beta}(V_{q+1}) \). For conciseness, we will write \( \Pr\{C_{\alpha,\beta}(V_q)\} \) for \( \Pr\{(x, x') \in C_{\alpha,\beta}(V_q)\} \).
Result in [17] implies \( V_{q+1} \subseteq B \left( \frac{\Pr\{C_{\alpha,\beta}(V_q)\}}{2\xi} \right) \) and
\[
\Pr\{C_{\alpha,\beta}(V_{q+1})\} \leq \Pr \left\{ C_{\alpha,\beta} \left( B_{\alpha,\beta} \left( \frac{\Pr\{C_{\alpha,\beta}(V_q)\}}{2\xi} \right) \right) \right\} \leq \xi \cdot \frac{\Pr\{C_{\alpha,\beta}(V_q)\}}{2\xi} = \frac{\Pr\{C_{\alpha,\beta}(V_q)\}}{2},
\]
where the first inequality is by the definition of \( \xi \). This result means \( \Pr\{C_{\alpha,\beta}(V_q)\} \) is halved after each round of labeling. Therefore, after \( Q := \log_2 \frac{1}{\varepsilon} \) rounds of labeling,
\[
\Delta_{\alpha,\beta}(h) \leq \Pr\{C_{\alpha,\beta}(V_Q)\} \leq \varepsilon,
\]
with probability at least \( 1 - Q\delta' \); where the left inequality is by definition. By then, the total number of labeled instances is \( \log_2 \frac{1}{\varepsilon} \cdot \frac{1}{4\xi^2} \left( 32c/\beta + \sqrt{\frac{1}{2} \log \frac{1}{\delta'}} \right) \). Setting \( \delta = Q\delta' \) and plugging \( \delta' = \delta/Q \) in completes the proof.

**Lemma 0.4 (Lemma 5.1).** Fix any \( \alpha, \beta > 0 \). We have \( \Delta_{\alpha,\beta}(h; S) \leq \hat{\Delta}_{\alpha,\beta}(h; S) \) for any \( h \in S \) and sample \( S \).

**Proof.** Since \( \mathbb{I}_{x \geq t} \leq \frac{t}{t} \) for any \( x, t \geq 0 \), we have
\[
\mathbb{I}\{d(x_i, x_j) \leq \alpha, |h(x_i) - h(x_j)| \geq \beta\} = \mathbb{I}\{d(x_i, x_j) \leq \alpha\} \cdot \mathbb{I}\{|h(x_i) - h(x_j)|^2 \geq \beta^2\}
\leq \frac{1}{\beta^2} \cdot \mathbb{I}\{d(x_i, x_j) \leq \alpha\} \cdot |h(x_i) - h(x_j)|^2
\leq \frac{1}{\beta^2} \cdot M_{ij} \cdot |h(x_i) - h(x_j)|^2.
\]
(20)
Plugging this back to (6) proves the lemma.

**References**


