# 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* $F : X \times X \to \mathbb{R}$ *be a hypothesis class induced from* $H$ *such that* $\forall f \in F$, $f(x, x') = \tau_\beta^t(|h(x) - h(x')|)$ *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 $G : X \times X \to \mathbb{R}$ be the set of functions induced from $h$ and defined as $\forall g \in G$, $g(a, b) = h(a) - h(b)$. Let $abs$ be the absolute function. Then $f(a, b) = \tau_\beta^t \circ abs \circ g(a, b)$ and we can write, accordingly,

$$F = \tau_\beta^t \circ abs \circ G. \tag{1}$$

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

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

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 $abs$ 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

$$\begin{aligned}
\mathcal{R}_m(G) &= \mathbb{E}_{\{(a_i,b_i)\}}\mathbb{E}_\sigma \sup_{g \in G} \frac{1}{m}\sum_{i=1}^m \sigma_i g(a_i, b_i) \\
&= \mathbb{E}_{\{(a_i,b_i)\}}\mathbb{E}_\sigma \sup_{g \in G} \frac{1}{m}\sum_{i=1}^m \sigma_i [h(a_i) - h(b_i)] \\
&\leq \mathbb{E}_{\{(a_i,b_i)\}}\mathbb{E}_\sigma \sup_{g \in G} \frac{1}{m}\sum_{i=1}^m \sigma_i h(a_i) + \mathbb{E}_{\{(a_i,b_i)\}}\mathbb{E}_\sigma \sup_{g \in G} \frac{1}{m}\sum_{i=1}^m \sigma_i h(b_i) \\
&= 2 \cdot \mathbb{E}_{\{(a_i,b_i)\}}\mathbb{E}_\sigma \sup_{g \in G} \frac{1}{m}\sum_{i=1}^m \sigma_i h(x_i) \\
&= 2 \cdot \mathcal{R}_m(H),
\end{aligned} \tag{3}$$

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. □

**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* $\Delta_{\alpha,\beta+1/t}(h) \leq \varepsilon$ *with probability at least* $1 - \delta$ *if* $m \geq \frac{1}{\varepsilon^2}\left(16tc + \sqrt{\frac{1}{2}\log\frac{1}{\delta}}\right)$, *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

$$\tau_\beta(z) = \begin{cases} 1, & \text{if } z > \beta \\ 0, & \text{if } z \le \beta \end{cases}, \tag{4}$$

and

$$\tau_\beta^t(z) = \begin{cases} 1, & \text{if } z > \beta + \frac{1}{t} \\ t(z-\beta), & \text{if } \beta < z \le \beta + \frac{1}{t} \\ 0, & \text{if } z \le \beta \end{cases}. \tag{5}$$

By definition, we have

$$\tau_{\beta+\frac{1}{t}}(z) \le \tau_\beta^t(z) \le \tau_\beta(z). \tag{6}$$

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

$$S_\alpha = \{(a,b) \in S \mid d(a,b) \le \alpha\}. \tag{7}$$

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

$$\begin{aligned} \Delta_{\alpha,\beta}(h;S) &= \frac{1}{n^2} \sum_{i,,j=1}^{n} \mathbb{I}\{|h(x_i) - h(x_j)| > \beta,\ d(x_i,x_j) \le \alpha\} \\ &= \frac{m}{n^2} \cdot \frac{1}{m} \sum_{(a,b)\in S_\alpha} \mathbb{I}\{|h(a) - h(b)| > \beta\} \\ &= \frac{m}{n^2} \cdot \frac{1}{m} \sum_{(a,b)\in S_\alpha} \tau_\beta(|h(a) - h(b)|). \end{aligned} \tag{8}$$

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

$$\begin{aligned} \frac{1}{m} \sum_{(a,b)\in S_\alpha} \tau_\beta(|h(a) - h(b)|) &\ge \frac{1}{m} \sum_{(a,b)\in S_\alpha} \tau_\beta^t(|h(a) - h(b)|) \\ &\ge \mathbb{E}[\tau_\beta^t(|h(a) - h(b)|) \mid d(a,b) \le \alpha] - 2\mathcal{R}_m(F) - \sqrt{\frac{\log\frac{1}{\delta}}{2m}} \\ &\ge \mathbb{E}[\tau_{\beta+\frac{1}{t}}(|h(a) - h(b)|) \mid d(a,b) \le \alpha] - 16t\mathcal{R}_m(H) - \sqrt{\frac{\log\frac{1}{\delta}}{2m}} \\ &\ge \mathbb{E}[\tau_{\beta+\frac{1}{t}}(|h(a) - h(b)|) \mid d(a,b) \le \alpha] - \frac{1}{\sqrt{m}}\left(16tc + \sqrt{\frac{1}{2}\log\frac{1}{\delta}}\right). \end{aligned} \tag{9}$$

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

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

$$\mathbb{E}[\tau_{\beta+\frac{1}{t}}(|h(a) - h(b)|) \mid d(a,b) \le \alpha] \le \frac{1}{m}\left(16tc + \sqrt{\frac{1}{2}\log\frac{1}{\delta}}\right). \tag{10}$$

[1]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

$$\Delta_{\alpha,\beta+\frac{1}{t}}(h) \leq \mathbb{E}[\tau_{\beta+\frac{1}{t}}(|h(a)-h(b)|) \mid d(a,b) \leq \alpha], \tag{11}$$

because

$$\begin{aligned}
\Delta_{\alpha,\beta+\frac{1}{t}}(h) &= \int_{(a,b)\in X\times X} \mathbb{I}\{|h(a)-h(b)| > \beta + 1/t\} \cdot \mathbb{I}\{d(a,b) \leq \alpha\} \cdot p(a,b) \\
&\leq \int_{(a,b)\in X\times X} \mathbb{I}\{|h(a)-h(b)| > \beta + 1/t\} \cdot p(a,b) \\
&\leq \int_{(a,b)\in X\times X} \mathbb{I}\{|h(a)-h(b)| > \beta + 1/t\} \cdot p(a,b \mid d(a,b) \leq \alpha) \\
&= \mathbb{E}[\tau_{\beta+\frac{1}{t}}(|h(a)-h(b)|) \mid d(a,b) \leq \alpha].
\end{aligned} \tag{12}$$

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

$$m \geq \frac{1}{\varepsilon^2}\left(16tc + \sqrt{\frac{1}{2}\log\frac{1}{\delta}}\right). \tag{13}$$

The theorem is proved. □

**Theorem 0.3** (Theorem 4.2)**.** *Fix any $\alpha, \beta > 0$. Suppose $\mathcal{R}_m(H) \in O(1/\sqrt{m})$ 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 $O(\log\frac{1}{\varepsilon})$ 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

$$V_q = \{h \in H; \Delta_{\alpha,\beta}(h; S_q) = 0\}. \tag{14}$$

Consider labeling $m$ instances in round $q+1$. First, note that all labeled instances fall in $\mathcal{C}_{\alpha,\beta}(V_q)$ 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 $m \geq \frac{1}{4\xi^2}\left(32c/\beta + \sqrt{\frac{1}{2}\log\frac{1}{\delta'}}\right)$, with probability at least $1-\delta'$, any $h \in V_{q+1}$ satisfies

$$\Delta_{\alpha,\beta}(h) \leq 1/(2\xi). \tag{15}$$

Let $\&$ be logic 'AND' and define event

$$I_\alpha^\beta(x,x';h) := d(x,x') \leq \alpha \;\&\; |h(x)-h(x')| > \beta. \tag{16}$$

Then, with probability at least $1-\delta'$, any $h \in V_{q+1}$ satisfies

$$\begin{aligned}
\Pr\{I_\alpha^\beta(x,x';h)\} &= \Pr\{I_\alpha^\beta(x,x';h) \,\&\, (x,x') \in \mathcal{C}_{\alpha,\beta}(V_q)\} + \Pr\{I_\alpha^\beta(x,x';h) \,\&\, (x,x') \notin \mathcal{C}_{\alpha,\beta}(V_q)\} \\
&= \Pr\{I_\alpha^\beta(x,x';h) \,\&\, (x,x') \in \mathcal{C}_{\alpha,\beta}(V_q)\} \\
&= \Pr\{I_\alpha^\beta(x,x';h) \mid (x,x') \in \mathcal{C}_{\alpha,\beta}(V_q)\} \cdot \Pr\{(x,x') \in \mathcal{C}_{\alpha,\beta}(V_q)\} \\
&\leq \frac{\Pr\{(x,x') \in \mathcal{C}_{\alpha,\beta}(V_q)\}}{2\xi},
\end{aligned} \tag{17}$$

where the second equality is by the fact that $\Pr\{I_\alpha^\beta(x,x';h) \;\&\; (x,x') \notin \mathcal{C}_{\alpha,\beta}(V_q)\} \leq \Pr\{I_\alpha^\beta(x,x';h) \;\&\; (x,x') \notin \mathcal{C}_{\alpha,\beta}(V_{q+1})\} = 0$, and the inequality is by (15) conditioned on an additional fact that all labeled instances fall in $\mathcal{C}_{\alpha,\beta}(V_{q+1})$. For conciseness, we will write $\Pr\{\mathcal{C}_{\alpha,\beta}(V_q)\}$ for $\Pr\{(x,x') \in \mathcal{C}_{\alpha,\beta}(V_q)\}$ .

Result in (17) implies $V_{q+1} \subseteq \mathcal{B}\left(\frac{\Pr\{\mathcal{C}_{\alpha,\beta}(V_q)\}}{2\xi}\right)$ and

$$\Pr\{\mathcal{C}_{\alpha,\beta}(V_{q+1})\} \leq \Pr\left\{\mathcal{C}_{\alpha,\beta}\left(\mathcal{B}_{\alpha,\beta}\left(\frac{\Pr\{\mathcal{C}_{\alpha,\beta}(V_q)\}}{2\xi}\right)\right)\right\} \leq \xi \cdot \frac{\Pr\{\mathcal{C}_{\alpha,\beta}(V_q)\}}{2\xi} = \frac{\Pr\{\mathcal{C}_{\alpha,\beta}(V_q)\}}{2}, \tag{18}$$

where the first inequality is by the definition of $\xi$. This result means $\Pr\{\mathcal{C}_{\alpha,\beta}(V_q)\}$ is halved after each round of labeling. Therefore, after $Q := \log_2 \frac{1}{\varepsilon}$ rounds of labeling,

$$\Delta_{\alpha,\beta}(h) \leq \Pr\{\mathcal{C}_{\alpha,\beta}(V_Q)\} \leq \varepsilon, \tag{19}$$

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 $\Delta_{\alpha,\beta}(h; S) \leq \tilde{\Delta}_{\alpha,\beta}(h; S)$ for any $h \in S$ and sample $S$.*

*Proof.* Since $\mathbb{I}_{x \geq t} \leq \frac{x}{t}$ for any $x, t \geq 0$, we have

$$\begin{aligned} \mathbb{I}\{d(x_i, x_j) \leq \alpha,\, |h(x_i) - h(x_j)| \geq \beta\} &= \mathbb{I}\{d(x_i, x_j) \leq \alpha\} \cdot \mathbb{I}\{|h(x_i) - h(x_j)|^2 \geq \beta^2\} \\ &\leq \frac{1}{\beta^2} \cdot \mathbb{I}\{d(x_i, x_j) \leq \alpha\} \cdot |h(x_i) - h(x_j)|^2 \\ &= \frac{1}{\beta^2} \cdot M_{ij} \cdot |h(x_i) - h(x_j)|^2. \end{aligned} \tag{20}$$

Plugging this back to (6) proves the lemma. □

## References


Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Mehryar Mohri and Afshin Rostamizadeh. Rademacher complexity bounds for non-iid processes. *Advances in Neural Information Processing Systems*, 21, 2008.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

Gal Yona and Guy Rothblum. Probably approximately metric-fair learning. In *International Conference on Machine Learning*, pages 5680–5688. PMLR, 2018.