FAIRNESS VIA ADVERSARIAL ATTRIBUTE NEIGHBOURHOOD ROBUST LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Improving fairness between privileged and less-privileged sensitive attribute groups (e.g, race, gender) has attracted lots of attention. To enhance the model performs uniformly well in different sensitive attributes, we propose a principled Robust Adversarial Attribute Neighbourhood (RAAN) loss to debias the classification head and to promote a fairer representation distribution across different sensitive attribute groups. The key idea of RAAN is to mitigate the differences of biased representations between different sensitive attribute groups by assigning each sample an adversarial robust weight, which is defined on the representations of adversarial attribute neighbors, i.e, the samples from different protected groups. To provide efficient optimization algorithms, we cast the RAAN into a sum of coupled compositional functions and propose a stochastic adaptive (Adam-style) and non-adaptive (SGD-style) algorithm framework SCRAAN with provable theoretical guarantee. Extensive empirical studies on fairness-related benchmark datasets verify the effectiveness of the proposed method.

1 INTRODUCTION

With the excellent performance, machine learning methods have penetrated into many fields and brought impact into our daily lifes, such as the recommendation (Lin et al., 2022; Zhang, 2021), sentiment analysis (Kiritchenko & Mohammad, 2018; Adragna et al., 2020) and facial detection systems (Buolamwini & Gebru, 2018). Due to the existing bias and confounding factors in the training data (Fabbrizzi et al., 2022; Torralba & Efros, 2011), model predictions are often correlated with sensitive attributes, e.g, race, gender, which leads to undesirable outcomes. Hence, fairness concern has become an increasingly prominent issue. For example, the job recommendation system recommends lower wage jobs more likely to women than men (Zhang, 2021). Buolamwini & Gebru (2018) proposed an intersectional approach that quantitatively show that three commercial gender classifiers, proposed by Microsoft, IBM and Face++, have higher error rate for the darker-skinned populations.

To alleviate the effect of spurious correlations[1] between the sensitive attribute groups and prediction, many bias mitigation methods have been proposed to learn a debiased representation distribution at encoder level by taking the advantage of the adversarial learning (Wang et al., 2019; Wadsworth et al., 2018; Edwards & Storkey, 2015; Elazar & Goldberg, 2018), causal inference (Singh et al., 2020; Kim et al., 2019) and invariant risk minimization (Adragna et al., 2020; Arjovsky et al., 2019). Recently, in order to further improve the performance and reduce computational costs for large-scale data training, learning a classification head using the representation of pretrained models have been widely used for different tasks. Taking image classification for example, the downstream tasks are trained by finetuning the classification head of ImageNet pretrained ResNet (He et al., 2016) model (Qi et al., 2020b; Kang et al., 2019). However, the pretrained model may introduce the undesiarable bias for the down steaming tasks. Debiasing the encoder of pretrained models to have fairer representations by retraining is time-consuming and compuatational expensive. Hence, debiasing the classification head on biased representations is also of great importance.

In this paper, we raise two research questions: *Can we improve the fairness of the classification head on a biased representation space? Can we further debias the representation space?* We give affirmative answers by proposing a Robust Adversarial Attribute Neighborhood (RAAN) loss.

[1] misleading heuristics that work for most training examples but do not always hold.

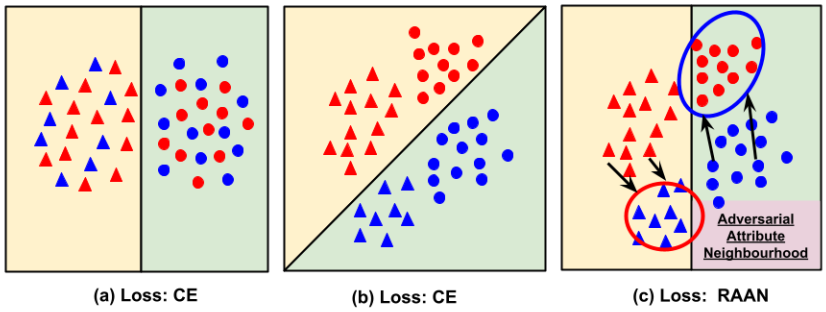

Figure 1: The influence of different protected group distributions on the classification head. The *colors* ({*red, blue*}) represent the sensitive attributes and *shapes* ({*triangle, circle*}) represent the ground truth class labels. Figures (a), (b) are optimized using vanilla CE loss, while the figure (c) is optimized using the proposed RAAN loss defined on the adversarial attributes neighborhood. The yellow and green background denotes the predicted classification space.

Our work is inspired by the RNF method (Du et al., 2021), which averages the representation of sample pairs from different protected groups to alleviate the undesirable correlation between fairness sensitive information and specific class labels. But unlike RNF, RAAN obtains fairness-promoting adversarial robust weights by exploring the AAN representation structure for each sample to mitigate the differences of biased sensitive attribute representations. To be more specific, the adversarial robust weight for each sample is the aggregation of the pairwise robust weights defined on the representation similarity between the sample and its AAN. Hence, the greater the representation similarity, the more uniform the distribution of protected groups in the representation space. Therefore, by promoting higher pairwise weights for larger similarity pairs, RAAN is able to mitigate the discrimination from the biased senstive attribute representations and promote a fairer classification head. When the representation is fixed, RAAN is also applicable to debiasing the classification head only.

We use a toy example of binary classification to express the advantages of RAAN over standard cross-entropy (CE) training on biased the sensitive attribute group distributions in Figure 1. Figure 1 (a) represents a uniform/fair distribution across different sensitive attributes while a biased distribution that the *red* attribute samples are more aggregated in the top left area than the *blue* attribute is depicted in Figure 1 (b), (c). Then with the vanilla CE training, Figure 1 (a) ends up with a fair classifier determined by the ground truth task labels (*shapes*) while a biased classification head determined by sensitive attributes (*colors*) is generated in Figure 1 (b). Instead, our RAAN method generates a fair classifier in Figure 1 (c), the same as a classifier learned from the Figure 1 (a) generated from a fair distribution. To this end, the main contributions of our work are summarized below:

- We propose a robust loss RAAN to debias the classification head by assigning adversarial robust weights defined on the top of biased representation space. When the representation is parameterized by trainable encoders such as convolutional layers in ResNets, RAAN is able to further debiase the representation distribution.
- We propose an efficient StochastiC algorithm framework for RAAN (SCRAAN), which includes the SGD-style and Adam-style updates with theoretical guarantee.
- Empirical studies on fairness-related datasets verify the supreme performance of the proposed SCRAAN on two fairness, Equalized Odd difference (ΔEO), Demographic Parity difference (ΔDP) and worst group accuracy.

2 Related Work

Bias Mitigation To address the social bias towards certain demographic groups in deep neural network (DNN) models (Lin et al., 2022; Zhang, 2021; Kiritchenko & Mohammad, 2018; Adragna et al., 2020; Buolamwini & Gebru, 2018), many efficient methods have been proposed to reduce the model discrimination (Wang et al., 2019; Wadsworth et al., 2018; Edwards & Storkey, 2015; Kim et al., 2019; Elazar & Goldberg, 2018; Singh et al., 2020; Zunino et al., 2021; Rieger et al., 2020; Liu & Avci, 2019; Kusner et al., 2017; Kilbertus et al., 2017; Cheng et al., 2021; Kang et al., 2019). Most methods in the above literature mainly focus on improving the fairness of representation. The authors

of (Wang et al., 2019; Wadsworth et al., 2018; Edwards & Storkey, 2015; Elazar & Goldberg, 2018) took the advantage of the adversarial training to reduce the group discrimination information. Rieger et al. (2020); Zunino et al. (2021) made use of the model explainability to remove subset features that incurs bias, while Singh et al. (2020); Kim et al. (2019) concentrated on the causal fairness features to get rid of undesirable bias correlation in the training. Bechavod & Ligett (2017) used the surrogate functions of fairness metric as the regularizer to penalizing unfairness. However, directly working on a biased representation to improves classification-head remains rare. Recently, the RNF method (Du et al., 2021) averages the representation of sample pairs from different protected groups in the biased representation space to remove the bias in the classification head. In this paper, we propose a principled RAAN objective that is able to debiasing both the representation distribution and classification head.

Robust Loss Several robust loss has been proposed to improve the model robustness for different tasks. The general cross entropy (GCE) loss was proposed to solve the noisy label problem which emphasizes more on the clean samples (Zhang & Sabuncu, 2018). For the data imbalanced problem, distributionally robust learning (DRO) (Qi et al., 2020b; Li et al., 2020; Sagawa et al., 2019) and class balance loss (Cui et al., 2019; Cao et al., 2019) use instance-level and class-level robust information from losses to pay more attention on underrepresented groups, respectively. Recently, Sagawa et al. (2019) shows that group DRO is able to prevent the models learning the specific spurious correlations. The above robust objective are defined on the loss space with the assistance of label information. Exploiting useful information from feature representation to further benefit the specific task training remains under-explored.

Invariant Risk Minimization (IRM) IRM (Arjovsky et al., 2019) is a novel paradigm to enhance model generalization in domain adaptation by learning the invariant feature representations of samples across different "domains" or "environments". By optimizing a practical version of IRM in the toxicity classification use case study, Adragna et al. (2020) shows the strength of IRM over ERM in improving the fairness of classifiers that are trained on biased data and tested on unbiased data. To elicit an invariant feature representation, IRM is casted into a constrained (bi-level) optimization problem where the classifier $\mathbf{w}_c$ is constrained on a optimal uncertainty set. Instead, the RAAN objective constrains the adversarial robust weights $\mathbf{p}$ for each sample in pairwise representation similarity space penalized by KL divergence as we show in section 3.2. When the embedding $\mathbf{z}$ is parameterized by trainable features $\mathbf{w}_f$, RAAN generates a more uniform representation space across different sensitive groups.

Stochastic Optimization Recently, several stochastic optimization technique has been leveraged to design efficient stochastic algorithms with provable theoretical convergence for the robust surrogate objectives, such as F-measure (Zhang et al., 2018b), average precision (AP) (Qi et al., 2021), and area under curves (AUC) (Liu et al., 2019; 2018; Yuan et al., 2021). In this paper, we cast the fairness promoting RAAN loss as a two-level stochastic coupled compositional function with a general formulation of $\mathbb{E}_\xi[f(\mathbb{E}_\zeta g(\mathbf{w};\zeta,\xi))]$, where ξ,ζ are independent and ξ has a finite support. By exploring the advanced stochastic compositional optimization technique (Wang et al., 2017; Qi et al., 2020a), a stochastic algorithm SCRANN with both SGD-style and Adam-style updates is proposed to solve the RAAN with provable convergence.

3 Robust Adversarial Attribute Neighbourhood (RAAN) Loss

3.1 Notations

We first introduce some notations in this subsection. The collected data is denoted by $\mathcal{D} = \{\mathbf{d}\}_{i=1}^n = \{(\mathbf{x}_i, y_i, a_i)\}_{i=1}^n$, where $\mathbf{x}_i \in \mathcal{X}$ is the feature, $y_i \in \mathcal{Y}$ is the label, $a_i \in \mathcal{A}$ is the corresponding attribute (e.g., race, gender), and n is the number of samples. Then we can divide the data into different subsets based on labels and attributes. For any label $c \in \mathcal{Y}$ and attribute $a \in \mathcal{A}$, we denote $\underline{\mathcal{D}_a^c = \{(\mathbf{x}_i, y_i, a_i) | a_i = a \wedge y_i = c\}_{i=1}^n}$ and $\underline{\mathcal{D}^c = \{(\mathbf{x}_i, y_i, a_i) | y_i = c\}_{i=1}^n}$. Then we have $\mathcal{D}^c = \cup\{\mathcal{D}_a^c\}_{a=1}^{|A|}$. Given a deep neural network, the model weights $\mathbf{w}$ can be decomposed into two parts, the Feature presentation parameters $\mathbf{w}_f$ and the Classification head parameters $\mathbf{w}_c$, i.e, $\mathbf{w} = [\mathbf{w}_f, \mathbf{w}_c]$. For example, $\mathbf{w}_f$ and $\mathbf{w}_c$ are mapped into the convolutions layers and fully connected layers in ResNets, respectively. Let $\mathbf{z}_i(\mathbf{w}_f) = F(\mathbf{w}_f, \mathbf{x}_i)/\|F(\mathbf{w}_f, \mathbf{x}_i)\|$ denotes the embedding representations of the sample $\mathbf{d}_i$. $H(\mathbf{w}_c, F(\mathbf{w}_f, \mathbf{x}_i))$ represents the output of the classification head.

The key idea of RAAN is to assigns an fairness-promoting adversarial robust weight for each sample by exploring the AAN representation structure to improve the fairness accross different sensitive attributes. We denote adversarial robust weight as $p_i^{\text{AAN}}, \forall\ \mathbf{d}_i \sim \mathcal{D}$. $p_i^{\text{AAN}}$is an aggregation of the pairwise weights between the sample $\mathbf{d}_i = (\mathbf{x}_i, y_i = c, a_i = a)$ and its Adversarial Attribute Neighbours (AAN), i.e, the samples from the same class but with different attributes, $\mathcal{P}_a^c = \mathcal{D}^c \backslash \mathcal{D}_a^c$. The AAN of sample $\mathbf{d}_i$ is represented as $\mathcal{P}_i = \mathcal{P}_a^c$. For example, we consider a binary attributes $\{male, female\}$ and a sample belonging to the $male$ protected group, then its AAN is the collection of the $female$ attribute samples with the same class label $y_i = c$. Next, we denote the pairwise robust weights between the sample $\mathbf{d}_i$ and $\mathbf{d}_j \in \mathcal{P}_i$ in the representation space as p_{ij}^{AAN}. And we abuse the notations by using $\mathbf{p}_i^{\text{AAN}} = [p_{i1}^{\text{AAN}}, \cdots, p_{ij}^{\text{AAN}}, \cdots] \in \mathbb{R}^{|\mathcal{P}_i|}$ to represents the pairwise robust weights vector defined in $\mathcal{P}_i$, the AAN of $\mathbf{d}_i$.

3.2 RAAN Objective

To explore the AAN representation structure and obtain the pairwise robust weights, we define the following robust constrained objective for $\forall \mathbf{d}_i \sim \mathcal{D}$,

$$\ell_i^{\text{AAN}} = \sum_{j \in \mathcal{P}_i} p_{ij}^{\text{AAN}} \ell(\mathbf{w}; \mathbf{x}_j, y_j, a_j) \tag{1}$$

$$\text{s.t.} \quad \max_{\mathbf{p}_i^{\text{AAN}} \in \Delta^{|\mathcal{P}_i|}} \sum_{j \in \mathcal{P}_i} p_{ij}^{\text{AAN}} \mathbf{z}_i(\mathbf{w}_f)^\top \mathbf{z}_j(\mathbf{w}_f) - \tau \text{KL}\left(\mathbf{p}_i^{\text{AAN}}, \frac{\mathbf{1}}{|\mathcal{P}_i|}\right),\ \mathbf{1} \in \mathbb{R}^{|\mathcal{P}_i|} \tag{2}$$

where Δ is a simplex that $\sum_{j=1}^{|\mathcal{P}_i|} p_{ij} = 1$. The robust loss (1) is a weighted average combination of the AAN loss. The robust constraint (2) is defined in the pairwise representation similarity between the sample i and its AAN penalized by the KL divergence regularizer, which has been extensively studied in distributionally robust learning objective (DRO) to improve the robustness of the model in the loss space (Qi et al., 2020b). Here, we adopt the DRO with KL divergence to the representation space to generate a uniform distribution across different sensitive attributes.

Controlled by the hyperparameter τ, the close form solution of $\mathbf{p}_i^{\text{AAN}}$ in (2) guarantees that the larger the pairwise similarity $\mathbf{z}_i^\top(\mathbf{w}_f)\mathbf{z}_j(\mathbf{w}_f)$ is, the higher the p_{ij}^{AAN} will be. When $\tau = 0$, the close form solution of (2)is 1 for the pair with the largest similarity and 0 on others. When $\tau > 0$, due to the strong convexity in terms of $\mathbf{p}_i^{\text{AAN}}$, the close form solution of (2) for each pair weight between $\mathbf{d}_i$ and $\mathbf{d}_j \in \mathcal{P}_i$ is:

$$p_{ij}^{\text{AAN}} = \frac{\exp(\frac{\mathbf{z}_i^\top(\mathbf{w}_f)\mathbf{z}_j(\mathbf{w}_f)}{\tau})}{\sum\limits_{k \in \mathcal{P}_i} \exp(\frac{\mathbf{z}_i^\top(\mathbf{w}_f)\mathbf{z}_k(\mathbf{w}_f)}{\tau})}. \tag{3}$$

Hence the larger the τ is, the more uniform of $\mathbf{p}_i^{\text{AAN}}$ will be. And it is apparent to see that the robust objective generates equal weights for every pair such that $p_{ij}^{\text{AAN}} = \frac{1}{|\mathcal{P}_i|}$ for every $\mathbf{d}_j \in \mathcal{P}_i$ when τ approaches to the infinity in Eqn (3). When we have a fair representation, the embeddings of different protected groups are uniform distributed in the representation space. The vanilla average loss training is good enough to have a fair classification head, which equals to RAAN with τ goes to infinity. When we have biased representations, we use a smaller τ to emphasize on the similar representations that shared invariant feature from two different protected groups to reduce the bias introduced from difference of the two protected group distributions.

To this end, after having the close form solution for every pairwise robust weights p_{ij}^{AAN} (3) in ℓ_i^{AAN} (1) given an arbitrary sample $i \sim \mathcal{D}$, the overall RAAN objective is defined as:

$$\text{RAAN}(\mathbf{w}) := \frac{1}{C}\sum_{c=1}^{C}\frac{1}{A}\sum_{a=1}^{A}\frac{1}{|\mathcal{D}_a^c|}\sum_{i=1}^{|\mathcal{D}_a^c|}\ell_i^{\text{AAN}} = \frac{1}{AC}\sum_{j=1}^{n} p_j^{\text{AAN}}\ell(\mathbf{w}; \mathbf{x}_j, y_j, a_j), \tag{4}$$

where $C = |\mathcal{Y}|$, $A = |\mathcal{A}|$, $\ell_i^{\text{AAN}} = \sum\limits_{j \in \mathcal{P}_i} p_{ij}^{\text{AAN}} \ell(\mathbf{w}; \mathbf{x}_j, y_j, a_j)$ is defined in (1) and p_{ij}^{AAN} is defined in (3), and $p_j^{\text{ANN}} = \frac{1}{|\mathcal{P}_j|}\sum\limits_{i \in \mathcal{P}_j} \frac{\exp(\frac{\mathbf{z}_i^\top(\mathbf{w}_f)\mathbf{z}_j(\mathbf{w}_f)}{\tau})}{\sum\limits_{k \in \mathcal{P}_i}\exp(\frac{\mathbf{z}_i^\top(\mathbf{w}_f)\mathbf{z}_k(\mathbf{w}_f)}{\tau})}$ is obtained by we aggregating all the pairwise

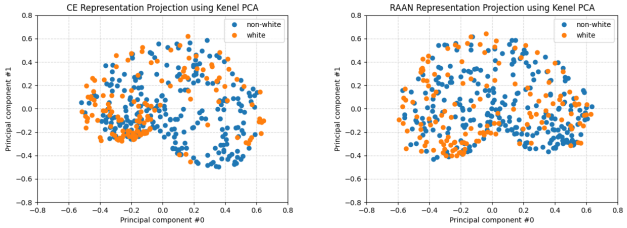

Figure 2: Improvement of Representation Fairness

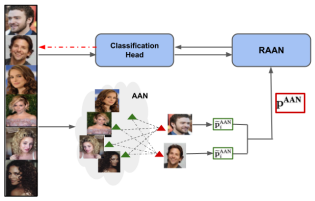

Figure 3: Overview of RAAN

robust weights. Hence, the adversarial robust weights p_j^{AAN} for each sample $\mathbf{d}_j \in \mathcal{D}$ encodes the intrinsic representation neighbourhood structure between the sample and its AAN samples $\mathbf{d}_i \in \mathcal{P}_j$ (the numerator) and normalized by the similarity pairs from the same protected groups $\mathbf{d}_k \in \mathcal{P}_i$ (the denominator). Due to the limitation of space, we put the second equality derivation of equation (4) in Appendix.

3.3 Representation Learning Robust Adversarial Attribute Neighbourhood (RL-RAAN)

AANs are defined over the encoder representation outputs, $\mathbf{z}(\mathbf{w}_f)$. By default, the RAAN is designed to promote a fairer classification head with a fixed representation encoder, i.e, $\mathbf{w}_f$ (recall that $\mathbf{w} = [\mathbf{w}_f, \mathbf{w}_c]$) is not trainable in Eqn (4). Here, we extend the RAAN to the Representation Learnining RAAN (RL-RAAN) by parameterizing the AANs with trainable encoder parameters, i.e, $\mathbf{w}_f$ is trainable. The red dashed arrow in Figure 3 represents the optional gradient backwards depending on whether $\mathbf{w}_f$ is trainable. Hence, RAAN optimizes the $\mathbf{w}_c$ while RL-RAAN jointly optimizes $[\mathbf{w}_f, \mathbf{w}_c]$. To design efficient stochastic algorithms, RL-RAAN requires more sophisticated stochastic estimators than RAAN, which we will discuss later in Section 4.

Here, we show that RL-RAAN is able to generate a more uniform representation distribution for different sensitive groups in Figure 2. To achieve this, we visualize the representation distribution of vanilla CE and RL-RAAN methods at the end of training using Kernel-PCA dimensionality reduction method with radial basis function (rbf) kernel. The left plot is the representation distribution learned using standard vanilla CE training and the right plot is the representation distribution at the end of RL-RAAN training. It is clear to see that *white*-sensitive attribute samples are more clustered in the upper left corner in CE while both the *white* and *non-white* sensitive attributes samples are both uniformly distributed in the representation space.

4 Stochastic Compositional optimization for RAAN

In this section, we provide a general Stochastic Compositional optimization algorithm framework for RAAN (SCRAAN). The SCRAAN applies to both RAAN and RL-RAAN objective. We first show that (RL)-RAAN is a **two-level stochastic coupled compositional function** and then design stochastic algorithms under the framework of stochastic gradient descent, SGD and Adam (Kingma & Ba, 2014) updates with theorectical guarantee in Algorithm 1.

Let $\mathbb{I}(c)$ denotes the indicator function that equals to 1 when c is true and equals to 0 otherwise. $g(\mathbf{w}; \mathbf{x}_i, \mathbf{x}_j) = [g_1(\mathbf{w}; \mathbf{x}_i, \mathbf{x}_j), g_2(\mathbf{w}; \mathbf{x}_i, \mathbf{x}_j)]^\top = \frac{n}{AC}[\exp(\frac{\mathbf{z}_i(\mathbf{w}_f)^\top \mathbf{z}_j(\mathbf{w}_f)}{\tau})\ell_j(\mathbf{w}; \mathbf{x}_j, c_j, a_j)\mathbb{I}(\mathbf{x}_j \in \mathcal{P}_i), \exp(\frac{\mathbf{z}_i(\mathbf{w}_f)^\top \mathbf{z}_j(\mathbf{w}_f)}{\tau})\mathbb{I}(\mathbf{x}_j \in \mathcal{P}_i)]^\top : \mathbb{R}^d \to \mathbb{R}^2$, $g_{\mathbf{x}_i}(\mathbf{w}) = \mathbb{E}[g(\mathbf{w}; \mathbf{x}_i)] = \mathbb{E}_{\mathbf{x}_j \in \mathcal{P}_i}[g(\mathbf{w}; \mathbf{x}_i, \mathbf{x}_j)]$, $f(g) = \frac{g_1}{g_2}, g = [g_1, g_2]$ and $f(s) = \frac{s_1}{s_2} : \mathbb{R}^2 \to \mathbb{R}$. Then the (RL)-RAAN objectiv (4) can be written as

$$\mathrm{R}(\mathbf{w}) = \frac{1}{n} \sum_{\mathbf{x}_i \in \mathcal{D}} f(g_{\mathbf{x}_i}(\mathbf{w})) = \mathbb{E}_{\mathbf{x}_i \in \mathcal{D}}[f(g_{\mathbf{x}_i}(\mathbf{w}))] \tag{5}$$

where g denotes the inner objective and f denotes the outer objective. The equivalence between (4) and (5) is shown in Appendix 6.5. In the following, we use $\widetilde{\mathbf{w}}$ to unify the trainable parameter notation of RAAN and RL-RAAN. Recall that $\mathbf{w} = [\mathbf{w}_f, \mathbf{w}_c]$, $\widetilde{\mathbf{w}} = \mathbf{w}_c$ when $R(\mathbf{w})$ represents the RAAN, and $\widetilde{\mathbf{w}} = \mathbf{w}$ when $R(\mathbf{w})$ represents RL-RAAN. Then according to the chain rule, the gradient of $R(\mathbf{w})$ is

Algorithm 1: SCRAAN

1: **Input:** Initialize $\mathbf{w}^1 = [\mathbf{w}_f^1, \mathbf{w}_c^1]$.
2: **while** first stage **do**
3: Train the whole model $\mathbf{w}$ with standard CE loss
4: **end while**
5: **while** second stage **do**
6: **for** $t = 1, \ldots, T-1$ **do**
7: Draw a batch samples $\{(\mathbf{x}_i, y_i, a_i)\}_{i=1}^B$
8: $\mathbf{u} = \text{UG}(\mathcal{B}, \mathbf{u}, \mathbf{w}_t, u_0)$
9: Compute (biased) Stochastic Gradient Estimator $G(\mathbf{w}_t)$ by Equation (9)
10: Update $\mathbf{w}_{t+1}$ with a SGD-style method or by a Adam-style method

$$\mathbf{w}_{t+1} = \text{UW}(\mathbf{w}_t, G(\mathbf{w}_t))$$

11: **end for**
12: **end while**
13: **Return:** $\mathbf{w}_R$, R is a index sampled from $1 \cdots T$.

Algorithm 2: UG($\mathcal{B}, \mathbf{u}, \mathbf{w}_t, \gamma, u_0$)

1: **for** each $\mathbf{x}_i \in \mathcal{B}$ **do**
2: Construct $\widehat{\mathcal{P}}_i = \mathcal{P}_i \cap \mathcal{B}$ and compute $[\widehat{g}_{\mathbf{x}_i}(\mathbf{w}_t)]_1, [\widehat{g}_{\mathbf{x}_i}(\mathbf{w}_t)]_2$ by Equation (8)
3: Compute $\mathbf{u}_{\mathbf{x}_i}^1 = (1-\gamma)\mathbf{u}_{\mathbf{x}_i}^1 + \gamma[\widehat{g}_{\mathbf{x}_i}(\mathbf{w}_t)]_1$
$\mathbf{u}_{\mathbf{x}_i}^2 = \max((1-\gamma)\mathbf{u}_{\mathbf{x}_i}^2 + \gamma[\widehat{g}_{\mathbf{x}_i}(\mathbf{w}_t)]_2, u_0)$
4: **end for**
5: **Return** $\mathbf{u}$

Algorithm 3: UW($\mathbf{w}_t, G(\mathbf{w}_t)$)

1: Option 1: SGD-style update (paras: α)
$\mathbf{w}_{t+1} = \mathbf{w}_t - \alpha G(\mathbf{w}_t)$
2: Option 2: Adam-style update (paras: $\alpha, \epsilon, \eta_1, \eta_2$) $h_{t+1} = \eta_1 h_t + (1-\eta_1)G(\mathbf{w}_t)$
$v_{t+1} = \eta_2 \hat{v}_t + (1-\eta_2)(G(\mathbf{w}_t))^2$
$\mathbf{w}_{t+1} = \mathbf{w}_t - \alpha \frac{h_{t+1}}{\sqrt{\epsilon + \hat{v}_{t+1}}}$
where $\hat{v}_t = v_t$ (Adam) or $\hat{v}_t = \max(\hat{v}_{t-1}, v_t)$ (AMSGrad)
3: **Return:** $\mathbf{w}_{t+1}$

$$\nabla_{\widetilde{\mathbf{w}}} R(\mathbf{w}) = \frac{1}{n}\sum_{\mathbf{x}_i \in \mathcal{D}} \nabla_{\widetilde{\mathbf{w}}} g_{\mathbf{x}_i}(\mathbf{w})^\top \nabla f(g_{\mathbf{x}_i}(\mathbf{w})) = \frac{1}{n}\sum_{\mathbf{x}_i \in \mathcal{D}} ([\nabla_{\widetilde{\mathbf{w}}} g_{\mathbf{x}_i}(\mathbf{w})]_1^\top, [\nabla_{\widetilde{\mathbf{w}}} g_{\mathbf{x}_i}(\mathbf{w})]_2^\top) \begin{pmatrix} \frac{1}{[g_{\mathbf{x}_i}(\mathbf{w})]_2} \\ -\frac{[g_{\mathbf{x}_i}(\mathbf{w})]_1}{[g_{\mathbf{x}_i}(\mathbf{w})]_2^2} \end{pmatrix} \tag{6}$$

In Algorithms 1, we approximate the gradients of $\nabla R(\mathbf{w})$ with the stochastic estimators. Let $\mathcal{B}$ denotes a B sample set randomly generated from $\mathcal{D}$. For each sample $\mathbf{x}_i \in \mathcal{B}$, we approximate the $\nabla_{\widetilde{\mathbf{w}}} g_{\mathbf{x}_i}(\mathbf{w})$ using the stochastic gradient on the current batch, $\nabla_{\widetilde{\mathbf{w}}} \widehat{g}_{\mathbf{x}_i}(\mathbf{w})$, i.e, $([\nabla_{\widetilde{\mathbf{w}}} \widehat{g}_{\mathbf{x}_i}(\mathbf{w})]_1^\top, [\nabla_{\widetilde{\mathbf{w}}} \widehat{g}_{\mathbf{x}_i}(\mathbf{w})]_2^\top)$. Denotes the stochastic AAN samples in current batch $\mathcal{B}$ for the sample i is denoted as $\widehat{\mathcal{P}}_i = \mathcal{P}_i \cap \mathcal{B}$ and $\exp_{ij}^\tau(\mathbf{w}_f^t) = \exp(\frac{\mathbf{z}_i(\mathbf{w}_f^t)^\top \mathbf{z}_j(\mathbf{w}_f^t)}{\tau})$, then stochastic estimators for (RL)-RAAN are represented as:

$$[\nabla_{\widetilde{\mathbf{w}}} \widehat{g}_{\mathbf{x}_i}(\mathbf{w}_t)]_1^\top = \begin{cases} \frac{n}{AC}\frac{1}{|\widehat{\mathcal{P}}_i|}\sum_{\mathbf{x}_j \in \widehat{\mathcal{P}}_i} \exp_{ij}^\tau(\mathbf{w}_f^t) \nabla_{\mathbf{w}_c} \ell_j(\mathbf{w}_t)^\top & \text{RAAN} \\ \frac{n}{AC}\frac{1}{|\widehat{\mathcal{P}}_i|}\sum_{\mathbf{x}_j \in \widehat{\mathcal{P}}_i} \exp_{ij}^\tau(\mathbf{w}_f^t)(\nabla_{\mathbf{w}} \ell_j(\mathbf{w}_t)^\top + (\mathbf{z}_i(\mathbf{w}_f^t) + \mathbf{z}_j(\mathbf{w}_f^t))^\top \ell_j(\mathbf{w}_t)) & \text{RL-RAAN} \end{cases} \tag{7}$$

$[\nabla_{\widetilde{\mathbf{w}}} \widehat{g}_{\mathbf{x}_i}(\mathbf{w})]_2^\top$ is a $\mathbf{0}$ vector for RAAN and the dimension of $\mathbf{0}$ equals to the dimension of $\mathbf{w}_c$. For RL-RAAN, equals to $[\nabla_{\widetilde{\mathbf{w}}} \widehat{g}_{\mathbf{x}_i}(\mathbf{w})]_2^\top = \frac{n}{AC|\widehat{\mathcal{P}}_i|}\sum_{\mathbf{x}_j \in \widehat{\mathcal{P}}_i} \exp_{ij}^\tau(\mathbf{w}_f)(\mathbf{z}_i(\mathbf{w}_f^t) + \mathbf{z}_j(\mathbf{w}_f^t))$.

To estimate $g_{\mathbf{x}_i}(\mathbf{w})$, however, the stochastic objective $\widehat{g}_{\mathbf{x}_i}(\mathbf{w})$ is not enough to control the approximation error such that the convergence of Algorithm 1 can be guaranteed. We borrow a technique from the stochastic compositional optimization literature (Wang et al., 2017) by using a moving average estimator to estimate $g_{\mathbf{x}_i}(\mathbf{w})$ for all samples. We maintain a matrix $\mathbf{u} = [\mathbf{u}_1, \mathbf{u}_2]$ and each of a column is indexed by a sample $\mathbf{x}_i \sim \mathcal{D}$ corresponding to the the moving average stochastic estimator of $g_{\mathbf{x}_i}(\mathbf{w})$. The Step 3 of Algorithm 2 describes the updates of $\mathbf{u}$, in which u_0 is a small constant to address the numeric issue that does not influence the convergence analysis and the stochastic estimator $\widehat{g}_{\mathbf{x}_i}(\mathbf{w})$ for sample i is

$$[\widehat{g}_{\mathbf{x}_i}(\mathbf{w}_t)]_1 = \frac{n}{AC}\frac{1}{|\widehat{\mathcal{P}}_i|}\sum_{j \in \widehat{\mathcal{P}}_i} \exp_{ij}^\tau(\mathbf{w}_f^t)\ell_j(\mathbf{w}_t), [\widehat{g}_{\mathbf{x}_i}(\mathbf{w}_t)]_2 = \frac{n}{AC}\frac{1}{|\widehat{\mathcal{P}}_i|}\sum_{j \in \widehat{\mathcal{P}}_i} \exp_{ij}^\tau(\mathbf{w}_f^t) \tag{8}$$

To sum up, the overall stochastic estimator $G(\mathbf{w})$ for $\nabla R(\mathbf{w})$ in a batch where the stochastic inner objective gradient estimator for (RL)-RAAN:

$$G(\mathbf{w}) = \frac{1}{B}\sum_{i=1}^B \nabla_{\widetilde{\mathbf{w}}} \hat{g}_{\mathbf{x}_i}(\mathbf{w})^\top \begin{pmatrix} \frac{1}{\mathbf{u}_{\mathbf{x}_i}^2} \\ -\frac{\mathbf{u}_{\mathbf{x}_i}^1}{[\mathbf{u}_{\mathbf{x}_i}^2]^2} \end{pmatrix} \tag{9}$$

Finally, we apply both the SGD-style and Adam-style updates for $\mathbf{w}$ in Algorithm 3 . Next we provide the theoretical analysis for SCRAAN.

Theorem 1 *Suppose Assumption 2 holds, $\forall\ t \in 1, \cdots, T$, and $T > n$, let the parameters be 1) $\alpha = \frac{1}{n^{2/5}T^{3/5}}, \gamma = \frac{n^{2/5}}{T^{2/5}}$ for the SGD updates; 2) $\eta_1 \leq \sqrt{\eta_2} \leq 1$, $\alpha = \frac{1}{n^{2/5}T^{3/5}}, \gamma = \frac{n^{2/5}}{T^{2/5}}$ for the AMSGrad updates. Then after running T iterations, SCRAAN with SGD-style updates or Adam-style update satisfies*

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\|\nabla R(\mathbf{w}_t)\|^2\right] \leq O\left(\frac{n^{2/5}}{T^{2/5}}\right),$$

where O suppresses constant numbers.

Remark: Even though RAAN and RL-RAAN enjoys the same iteration complexity in Theorem 1, the stochastic estimator $[\nabla_{\widetilde{\mathbf{w}}}\widehat{g}_{\mathbf{x}_i}(\mathbf{w})]_2^\top$ is $\mathbf{0}$ leads to a simpler optimization for RAAN such that we only need to maintain and update $\mathbf{u}_{\mathbf{x}_i}^2$ to calculate $G(\mathbf{w}_t) = 1/|\mathcal{B}|\sum_{i=1}^{|\mathcal{B}|}[\nabla_{\mathbf{w}_c} g_{\mathbf{x}_i}(\mathbf{w})]_1/\mathbf{u}_{\mathbf{x}_i}^2$. There are other more sophisticated optimizers, such as MOAP (Wang et al., 2022), BSGD (Hu et al., 2020), can also be applied to solve (4), which we leave as a future exploration direction. The derivation of Theorem 1 is provided in Appendix 6.6

5 Empirical studies

In this section, we conduct empirical studies on fourdatasets: Adult (Kohavi et al., 1996), Medical Expenditure (MEPS)(Cohen, 2003), CelebA (Liu et al., 2015), and Civil Comments[2] dataset in the NLP. We compare the proposed methods with: 1) bias mitigation methods: RNF (Du et al., 2021), Adversarial learning (Zhang et al., 2018a), regularization method (Bechavod & Ligett, 2017) 2) robust optimization methods: Empirical Risk Minimization (ERM) and Invariant Risk Minimization (IRM)(Adragna et al., 2020; Arjovsky et al., 2019).

Datasets: For the two benchmark tabular datasets, the Adult dataset used to predict whether a person's annual income higher than 50K while the goal of MEPS is to predict whether a patient could have a high utilization. For the CelebA image dataset, we want to predict whether a person has wavy hair or not. Civil Comments dataset is an NLP dataset aims to predict the binary toxicity label for online comments. Accordingly, the protected sensitive attribute is **gender** $\in \{female, male\}$ [3]on the Adult and CelebA datasets, and the protected sensitive attribute is **race** $\in \{white, nonwhite\}$ on MEPS. We consider four different types of demographic sensitive attributes for each comments belonging to {**Black, Muslim, LGBTQ, NeuroDiverse**} on the Civil Comments dataset. The training data size varies from 11362 in MEPS, 33120 in Adult to 194599 in CelebA. Civil Comments Dataset contains 2 million online news articles comments that are annotated by toxicity. Following the setting in (Adragna et al., 2020), subsets of 450,000 comments for each sensitive attribute are constructed.

Metrics: In the experiments, we compare two fairness metric equalized odd difference (ΔEO), demographic parity difference (ΔDP) between different methods given the same accuracy and worst group accuracy in terms of {*Class* $\times$ *Attribute*}. ΔDP measures the difference in probability of favorable outcomes between unprivileged and privileged groups $\Delta\text{DP} = (\text{PR}_0 - \text{PR}_1)$, where $\text{PR}_0 = p(\hat{y} = 1|a = 0)$, and $\text{PR}_1 = p(\hat{y} = 1|a = 1)$. ΔEO requires favorable outcomes to be independent of the protected class attribute a, conditioned on the ground truth label $\mathbf{y}$. $\Delta\text{EO} = (\text{TPR}_0 - \text{TPR}_1) + (\text{FPR}_0 - \text{FPR}_1)$, where $\text{TPR}_0 = p(\hat{y} = 1|a = 0, y = 1)$, $\text{TPR}_1 = p(\hat{y} = 1|a = 1, y = 1)$, $\text{FPR}_0 = p(\hat{y} = 1|a = 0, y = 0)$, and $\text{FPR}_1 = p(\hat{y} = 1|a = 1, y = 0)$.

5.1 Comparison with Bias Mitigation methods

Baselines In this section we compare RAAN and RL-RAAN optimized by Adam-style SCRAAN with baselines optimized by Adam optimizer on Adult, MEPS and CelebA datasets. Correspondingly, the experimental results of the SGD-style optimizer including SCRAAN and other baselines are provided in Appendix 6.3. Among the baselines, **Vanilla** refers to the standard CE training with cross entropy loss, **RNF** represents the representation neutralization method in (Du et al., 2021), **Adversarial** denotes the adversarial training method (Zhang et al., 2018a) that mitigates biases by

[2] `https://www.tensorflow.org/datasets/catalog/civil_comments`

[3] For the gender attribute, there are more than binary attributes. For example, it contains but not limited to female, male and transgender are included to name a few. Here, due to the limited size of the datasets, we only consider female and male attributes in this paper.

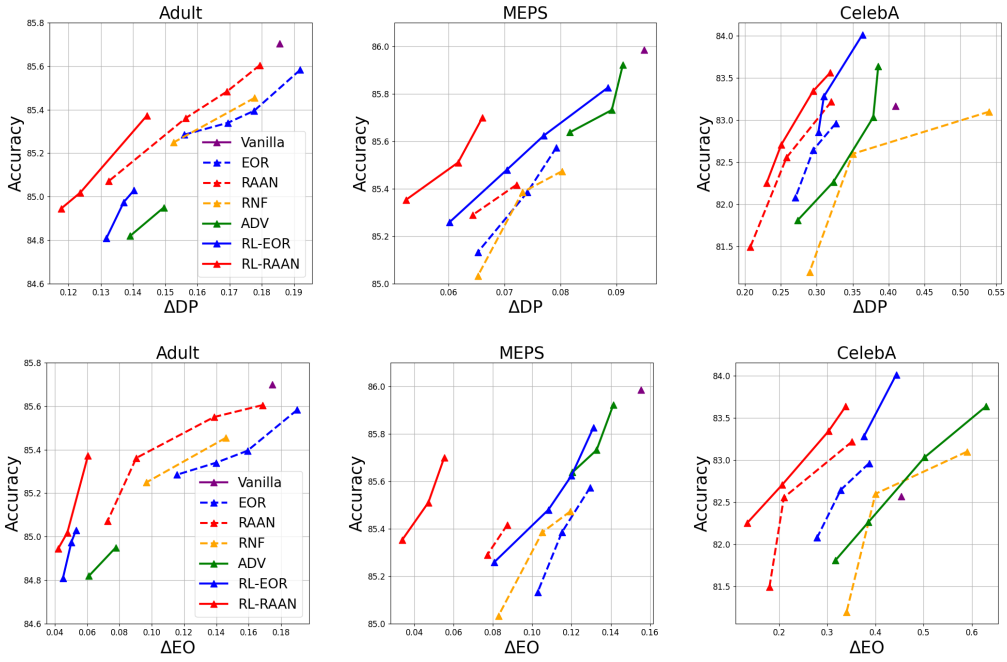

Figure 4: ΔDP and ΔEO experimental results of different methods optimized by Adam-style SCRAAN on Adult, MEPS and CELEBA, respectively. The results are reported over 5 runs.

simultaneously learning a predictor and an adversary, **(RL-)EOR** (Bechavod & Ligett, 2017) is a regularization method that uses a surrogate function of Δ EO as the regularizer. The comparison of baselines are described in Table 1. We report two version of experiments for the regularization method and our proposed method for debiasing the representation encoder and debiasing the classification head, i.e, RL-EOR and EOR, RL-RAAN and RAAN, respectively.

Table 1: Comparison of baseline methods

	Debiasing Representation Encoder	Debiasing Classification Head
Vanilla	×	×
Adversarial	✓	×
(RL-)EOR	✓	✓
RNF	×	✓
(RL-)RAAN	✓	✓

Models and Parameter settings Following the experimental setting in (Du et al., 2021), we train a three layer MLP for Adult and MEPS datasets and ResNet-18 for CelebA. The details of the MLP networks are provided in the Appendix 6.1. We adopt the two stage bias mitigation training scheme such that we apply the vanilla CE method in the first stage and then debias the representation encoder $\mathbf{w}_f$ or classification head $\mathbf{w}_c$ in the second stage. The representation encoder is fixed when we debias the classification head $\mathbf{w}_c$. We train 10 epochs per stage for MEPS and Adult datasets, and 5 epochs per stage for the CelebA. We report the final ΔOD, ΔDP and worst group accuracy in terms of $\{attribute \times label\}$ on the test data at the end of the training. The batch size of MEPS and Adult is 64 by default, and the batch size of celebA is 190. We use the Adam-style/SGD-style SCRAAN optimize RAAN, and Adam/SGD optimizer for other baselines. For all the methods, the learning rate α is tuned in $\{1e\text{-}2, 1e\text{-}3, 1e\text{-}4\}$ and $\tau \in \{0.1 : 0.2 : 2\}$. For the RAAN, we tune $\gamma \in \{0.1, 0.5, 0.9\}$. The regularizer hyperparameter of RNF α' is tuned in $\{0,\ 1e\text{-}4,\ 2e\text{-}4,\ 3e\text{-}4,\ 4e\text{-}4\}$. And the regularizer parameter in Adversarial and EOR is tuned $\in \{0.01 : 0.02 : 0.1\}$. The learning rate for the adversarial head α_{Adv} in Adversarial is tuned in $\{1e\text{-}2, 1e\text{-}3, 1e\text{-}4\}$.

Experimental results We present the Accuracy vs Δ DP, Accuracy vs Δ EO results for different methods in Figure 4. And the worst group accuracy are reported in Table 2. For the same accuracy, the smaller the value of Δ DP and Δ EO is, the better the method will be. It is worth to notice that we can not have a valid result for when replicating RNF method using SGD optimizer. By balancing between accuracy and Δ DP, Δ EO, RL-RAAN has the best results on all datasets. For the RAAN method, it has smallest Δ DP and Δ EO among the methods that debias on the classification head (the dashed lines). Besides the Vanilla method, EOR has the worst Δ DP and ΔEO in Adult, while Adversarial method has the worst Δ DP and ΔEO in MEPS and CelebA. When it comes to the worst

Table 2: Worst group accuracy over 5 independent runs.

	MEPS		Adult		CelebA	
Optimizer	SGD-style	Adam-style	SGD-style	Adam-style	SGD-style	Adam-style
Vanilla	27.61 ± 0.12	32.80 ± 0.15	55.49 ± 0.11	53.96 ± 0.14	41.32 ± 0.92	40.65 ± 0.85
EOR	27.82 ± 0.15	30.23 ± 0.15	51.01 ± 0.51	52.74 ± 0.21	46.02 ± 0.97	45.31 ± 1.58
RNF	-	31.91 ± 0.23	-	58.85 ± 0.41	-	50.23 ± 0.69
RAAN	**28.01** ± 0.17	**33.10** ± 0.10	**58.23** ± 0.13	**59.76** ± 0.31	**53.47** ± 1.24	**55.81** ± 1.12
Adversarial	28.31 ± 0.14	32.76 ± 0.13	54.72 ± 0.76	55.49 ±0.31	41.51 ± 0.98	40.15 ± 0.69
RL-EOR	29.52 ± 0.21	32.00 ± 0.23	59.15 ± 0.11	57.41 ± 0.41	44.61 ± 0.83	46.89 ± 1.51
RL-RAAN	**30.00** ± 0.45	**35.21** ± 0.33	**65.94** ± 0.86	**68.10** ± 0.39	**58.61** ± 1.01	**66.44** ± 0.72

group accuracy, RAAN and RL-RAAN achieve the best performance in debiasing the classification and representation encoder, respectively. In addition we provide ablation studies in the Apendx 6.2.

5.2 Comparison with Stochastic Optimization Methods

In this section, we compare RAAN with stochastic optimization methods including ERM, Group DRO, and IRM on the subsets of Civil Comments dataset. IRM and Group DRO have been proved to prevent models from learning prespecified spurious correlations (Adragna et al., 2020; Sagawa et al., 2019). We consider three different enviroments for training and testing (Adragna et al., 2020). We set the sample size to be the same for the three environments. For each environment, we have a balanced number of comments for each class and each attribute, i.e, half are non-toxic ($y = 0$) and half are toxic comments ($y = 0$). Similarly, for each sensitive demographic attribute in $\{$Black, Muslim, LGBTQ, NeuroDiverse$\}$, half of the comments are about the sensitive demographic attribute $(a = 1)$ and half are not $(a = 0)$. We define the label switching probability $p_e = p(a = z|y = 1 - z), \forall z \in \{0, 1\}$ to introduce the spurious correlations between the sensitive attributes and class labels and quantify the difference between different environments. The training datasets include two enviorments with $p_e = 0.1$ and 0.2, while $p_e = 0.9$ in testing data environment.

Baselines For the ERM, we optimizes vanilla CE using Adam optimizer. IRM optimize a practical variant objective for the linear invariant predictor, i.e, Equation (IRMv1), proposed in (Arjovsky et al., 2019) using Adam optimizer. Group DRO (Sagawa et al., 2019) aims to minimize the worst group accuracy. For proposed methods, we optimize the RAAN using the Adam-style SCRAAN. SCRAAN and Group DRO explicitly make use of the sensitive attributes information to construct AAN and calculate group loss, respectively.

Model and Parameter Settings We train a logistic regression with l2 regularization as the toxicity classification model (Adragna et al., 2020) by converting each comment into a sentence embedding representing its semantic content using a pre-trained Sentence-BERT model (Reimers & Gurevych, 2019). All the learning rates are finetuned using grid search between $\{0.0001, 0.01\}$. The hyper parameter of RAAN follows previous section. For the Group DRO the temperature parameter η is tuned in $\{1 : 0.2 : 2\}$. The hyperparemeter for IRM are tuned following (Adragna et al., 2020).

Experimental Results The experimental results are reported in Table 3. We can see that SCRAAN and Group DRO have a significant improvement over ERM and IRM on all three evaluation metric, which implies the effectiveness of sensitive attributes information to reduce model bias. When compared with Group DRO the SCRAAN and Group DRO, SCRAAN has comparable results in terms of group accuracy while performs better on ΔEO. This makes sense as the objective of Group DRO aims to minimize the worst group loss, while RAAN focuses on improving the fairness of different groups.

Table 3: Experimental results on the testing environments over 5 independent runs.

	Accuracy				Worst Group Accuracy				ΔEO			
Sens Att	ERM	IRM	Group DRO	SCRAAN	ERM	IRM	Group DRO	SCRAAN	ERM	IRM	Group DRO	SCRAAN
Black	47.04 ± 0.9	55.31 ± 1.2	67.32 ± 1.0	**71.29** ± 1.0	35.01 ± 0.7	45.01± 0.9	**64.91**± 1.0	64.23 ± 0.9	52.23 ± 3.4	30.90 ± 4.1	12.82 ± 2.7	**4.77** ± 2.1
Muslim	49.23 ± 0.9	59.08 ± 1.4	66.37± 1.2	**71.82** ± 1.0	36.92 ± 0.8	55.92 ± 1.7	62.45 ± 1.1	**62.51** ± 0.9	47.44 ± 2.1	25.93 ± 4.1	11.79 ± 2.2	**7.47** ± 1.9
NeuroDiv	65.18 ± 1.0	63.60 ± 1.3	68.03 ± 1.1	**68.17** ± 1.2	56.26 ± 1.0	45.75 ± 0.8	63.76 ±0.9	**64.06** ± 1.1	26.53 ± 1.7	26.26 ± 2.1	10.75 ± 1.6	**4.94** ± 0.9
LGBTQ	56.67 ± 1.1	61.99 ± 1.5	66.76 ± 1.3	**69.54** ± 1.1	42.58 ± 0.7	52.74 ± 1.2	63.10 ± 0.9	**67.31** ± 1.0	37.53 ± 2.1	25.00 ± 1.9	18.56 ± 2.1	**7.65** ± 1.3

6 Conclusion

In this paper, we propose a robust loss RAAN that is able to reduce the bias of the classification head and improve the fairness of representation encoder. Then an optimization framework SCRAAN has been developed for handling RAAN with provable theoretical convergence guarantee. Comprehensive studies on several fairness-related benchmark datasets verify the effectiveness of the proposed methods.

REFERENCES

Robert Adragna, Elliot Creager, David Madras, and Richard Zemel. Fairness and robustness in invariant learning: A case study in toxicity classification. *arXiv preprint arXiv:2011.06485*, 2020.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Yahav Bechavod and Katrina Ligett. Penalizing unfairness in binary classification. *arXiv preprint arXiv:1707.00044*, 2017.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pp. 77–91. PMLR, 2018.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019.

Tianyi Chen, Yuejiao Sun, and Wotao Yin. Solving stochastic compositional optimization is nearly as easy as solving stochastic optimization. *IEEE Transactions on Signal Processing*, 69:4937–4948, 2021.

Pengyu Cheng, Weituo Hao, Siyang Yuan, Shijing Si, and Lawrence Carin. Fairfil: Contrastive neural debiasing method for pretrained text encoders. *arXiv preprint arXiv:2103.06413*, 2021.

Steven B Cohen. Design strategies and innovations in the medical expenditure panel survey. *Medical care*, pp. III5–III12, 2003.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9268–9277, 2019.

Mengnan Du, Subhabrata Mukherjee, Guanchu Wang, Ruixiang Tang, Ahmed Awadallah, and Xia Hu. Fairness via representation neutralization. *Advances in Neural Information Processing Systems*, 34, 2021.

Harrison Edwards and Amos Storkey. Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897*, 2015.

Yanai Elazar and Yoav Goldberg. Adversarial removal of demographic attributes from text data. *arXiv preprint arXiv:1808.06640*, 2018.

Simone Fabbrizzi, Symeon Papadopoulos, Eirini Ntoutsi, and Ioannis Kompatsiaris. A survey on bias in visual datasets. *Computer Vision and Image Understanding*, pp. 103552, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Yifan Hu, Siqi Zhang, Xin Chen, and Niao He. Biased stochastic first-order methods for conditional stochastic optimization and applications in meta learning. *Advances in Neural Information Processing Systems*, 33:2759–2770, 2020.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*, 2019.

Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. *Advances in neural information processing systems*, 30, 2017.

Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9012–9020, 2019.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Svetlana Kiritchenko and Saif M Mohammad. Examining gender and race bias in two hundred sentiment analysis systems. *arXiv preprint arXiv:1805.04508*, 2018.

Ron Kohavi et al. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Kdd*, volume 96, pp. 202–207, 1996.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.

Tian Li, Ahmad Beirami, Maziar Sanjabi, and Virginia Smith. Tilted empirical risk minimization. *arXiv preprint arXiv:2007.01162*, 2020.

Shuo Lin, Jianling Wang, Ziwei Zhu, and James Caverlee. Quantifying and mitigating popularity bias in conversational recommender systems. *arXiv preprint arXiv:2208.03298*, 2022.

Frederick Liu and Besim Avci. Incorporating priors with feature attribution on text classification. *arXiv preprint arXiv:1906.08286*, 2019.

Mingrui Liu, Xiaoxuan Zhang, Zaiyi Chen, Xiaoyu Wang, and Tianbao Yang. Fast stochastic auc maximization with $o(1/n)$-convergence rate. In *International Conference on Machine Learning*, pp. 3189–3197. PMLR, 2018.

Mingrui Liu, Zhuoning Yuan, Yiming Ying, and Tianbao Yang. Stochastic auc maximization with deep neural networks. *arXiv preprint arXiv:1908.10831*, 2019.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

Qi Qi, Zhishuai Guo, Yi Xu, Rong Jin, and Tianbao Yang. A practical online method for distributionally deep robust optimization. 2020a.

Qi Qi, Yi Xu, Rong Jin, Wotao Yin, and Tianbao Yang. Attentional biased stochastic gradient for imbalanced classification. *arXiv preprint arXiv:2012.06951*, 2020b.

Qi Qi, Youzhi Luo, Zhao Xu, Shuiwang Ji, and Tianbao Yang. Stochastic optimization of areas under precision-recall curves with provable convergence. *Advances in Neural Information Processing Systems*, 34, 2021.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

Laura Rieger, Chandan Singh, William Murdoch, and Bin Yu. Interpretations are useful: penalizing explanations to align neural networks with prior knowledge. In *International conference on machine learning*, pp. 8116–8126. PMLR, 2020.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Krishna Kumar Singh, Dhruv Mahajan, Kristen Grauman, Yong Jae Lee, Matt Feiszli, and Deepti Ghadiyaram. Don't judge an object by its context: learning to overcome contextual bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11070–11078, 2020.

Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pp. 1521–1528. IEEE, 2011.

Christina Wadsworth, Francesca Vera, and Chris Piech. Achieving fairness through adversarial learning: an application to recidivism prediction. *arXiv preprint arXiv:1807.00199*, 2018.

Guanghui Wang, Ming Yang, Lijun Zhang, and Tianbao Yang. Momentum accelerates the convergence of stochastic auprc maximization. In *International Conference on Artificial Intelligence and Statistics*, pp. 3753–3771. PMLR, 2022.

Hao Wang, Berk Ustun, and Flavio Calmon. Repairing without retraining: Avoiding disparate impact with counterfactual distributions. In *International Conference on Machine Learning*, pp. 6618–6627. PMLR, 2019.

Mengdi Wang, Ethan X Fang, and Han Liu. Stochastic compositional gradient descent: algorithms for minimizing compositions of expected-value functions. *Mathematical Programming*, 161(1): 419–449, 2017.

Zhuoning Yuan, Yan Yan, Milan Sonka, and Tianbao Yang. Large-scale robust deep auc maximization: A new surrogate loss and empirical studies on medical image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3040–3049, 2021.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 335–340, 2018a.

Shuo Zhang. Measuring algorithmic bias in job recommender systems: An audit study approach. 2021.

Xiaoxuan Zhang, Mingrui Liu, Xun Zhou, and Tianbao Yang. Faster online learning of optimal threshold for consistent f-measure optimization. *Advances in Neural Information Processing Systems*, 31, 2018b.

Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31, 2018.

Andrea Zunino, Sarah Adel Bargal, Riccardo Volpi, Mehrnoosh Sameki, Jianming Zhang, Stan Sclaroff, Vittorio Murino, and Kate Saenko. Explainable deep classification models for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3233–3242, 2021.

APPENDIX

6.1 MLP NETWORK STRUCTURES

To gain the feature representations, we use a three layer MLP for both Adult and MEPS datasets. The input and hidden layers are following up with a ReLU activation layer and a 0.2 drop out layer, respectively. The input size is 120 for Adult dataset and 138 for the MEPS dataset. The hidden size is 50 for both datasets. After that, we use a two layer classification head with a ReLU and 0.2 drop out layer for the second stage training prediction.

6.2 ABLATION STUIDES OF SCRAAN

γ and τ are two key paramters for SCRAAN. τ is the key hyperparameter to control the pairwise robust weights aggregation for RAAN. γ is designed for the stability and theoretical guarantees of Algorithm 1. We provide ablation studies for the two parameters independently.

To analyze the robustness of Algorithm 1 in terms of γ, we report ΔEO, ΔDP given the accuracy 85.3 for the Adam-style SCRAAN and 84.95 for the SGD-style SCRAAN on Adult dataset in Figure 5 by varing $\gamma \in \{0.1 : 0.1 : 0.9\}$ and fixing $\tau = 0.9$. It is obvious to see that both SGD-style SCRAAN and Adam-style SCRAAN are robust enough to have valid fairness evaluations.

Similarly, for the parameter τ, we report ΔEO, ΔDP of Adam-style SCRAAN to achieve accuracy 85.3 on Adult dataset by varing $\tau = \{0.1 : 0.2 : 1.9\}$ with $\gamma = 0.5$. We can see that by hypertuning τ in a reason range, we are able to find a τ achieves lowest ΔEO and ΔDP at the same time.

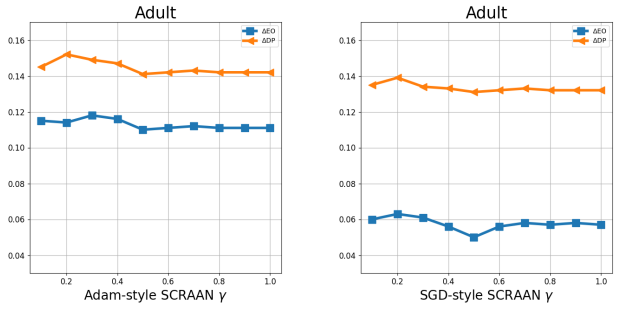

Figure 5: Robustness of γ.

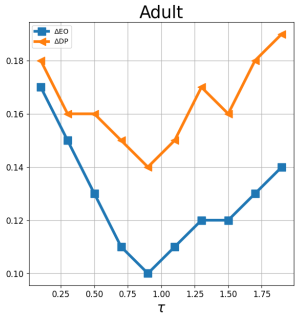

Figure 6: Influence of τ

6.3 MORE EXPERIMENTAL RESULTS OF SGD-STYLE SCRAAN

Here we provide the SGD-style SCRAAN experimental results on the Adult dataset. We can see that our methods are better than the baselines which is consistent with Adam-style SCRAAN.

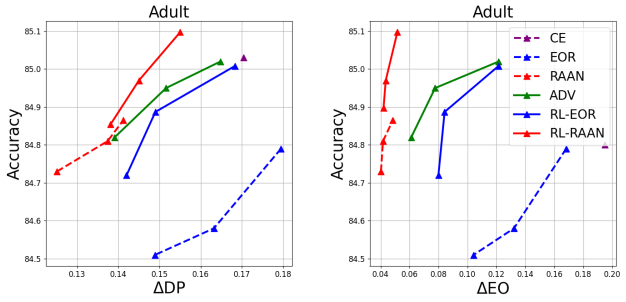

Figure 7: The ΔDP and ΔEO on the Adult dataset optimized by SGD-Style SCRAAN.

Table 4: We report class wise accuracy for $y \in \{0, 1\}$ and True Positive Rate and False Positive Rate for each group $a \in \{0, 1\}$, i.e, TPR_0, FPR_0, TPR_1 and FPR_1, for {Black, Muslim, LGBTQ, NeuroDiv} attributes respectively.

Black	Class 0	Class 1	TPR_0	FPR_0	TPR_1	FPR_1
ERM	41.23	52.69	48.77	4.74	92.98	64.99
IRM	62.57	47.85	45.40	7.37	73.68	40.88
Group DRO	66.59	67.79	67.12	19.47	77.19	35.01
SCRAAN	72.84	69.47	70.60	26.84	61.40	27.20
Muslim	Class 0	Class 1	TPR_0	FPR_0	TPR_1	FPR_1
ERM	42.49	55.5	52.37	8.11	92.05	63.31
IRM	59.13	58.26	57.05	12.43	77.27	44.07
Group DRO	64.29	67.58	67.90	19.46	73.30	37.55
SCRAAN	73.94	72.02	74.40	29.19	59.09	25.70
LGBTQ	Class 0	Class 1	TPR_0	FPR_0	TPR_1	FPR_1
ERM	50.33	61.68	88.12	54.48	60.43	0.071
IRM	52.19	70.33	88.13	51.07	70.26	18.93
Group DRO	65.59	63.36	66.71	12.43	79.38	36.90
SCRAAN	71.53	66.97	68.89	19.52	65.62	21.43
NeuroDiv	Class 0	Class 1	TPR_0	FPR_0	TPR_1	FPR_1
ERM	59.55	70.18	68.69	12.61	90.82	43.54
IRM	60.81	65.77	64.24	11.71	86.24	42.24
Group DRO	65.04	70.36	70.20	23.42	78.90	36.24
SCRAAN	66.31	69.37	69.80	27.93	72.48	34.33

6.4 More Experimental Results on Civil Comments

6.5 The Derivation of RAAN Objective, Equation (4), in Section 4

Given the pairwise weights between each sample $i \sim \mathcal{D}$ and its ANN, i.e, equation (1), (2). We have the following loss by averaging over all samples within the same protected groups, attributes and classes, wee have the following average neighbourhood robust loss.

Rewriting equation (4)

$$\frac{1}{C}\sum_{c=1}^{C}\frac{1}{|A|}\sum_{a=1}^{|A|}\frac{1}{|\mathcal{D}_a^c|}\sum_{i=1}^{|\mathcal{D}_a^c|}\underbrace{\sum_{j\in\mathcal{P}_i} p_{ij}^{\text{AAN}}\ell(\mathbf{w};\mathbf{x}_j,c,a_j)}_{\ell_i^{\text{AAN}}}$$

where $p_{ij}^{\text{AAN}} = \frac{\exp(\frac{\mathbf{z}_i^\top(\mathbf{w}_f)\mathbf{z}_j(\mathbf{w}_f)}{\tau})}{\sum_{k\in\mathcal{P}_i}\exp(\frac{\mathbf{z}_i^\top(\mathbf{w}_f)\mathbf{z}_k(\mathbf{w}_f)}{\tau})}$. To start with, p_{ij}^{AAN} is derived from the constraint robust pairwise objective in Equation (2),

$$\max_{\mathbf{p}_i^{\text{AAN}}\in\Delta^{|\mathcal{P}_i|}}\sum_{j\in\mathcal{P}_i} p_{ij}^{\text{AAN}}\mathbf{z}_i(\mathbf{w}_f)^\top\mathbf{z}_j(\mathbf{w}_f) - \tau\text{KL}(\mathbf{p}_i^{\text{AAN}}\|\frac{1}{|\mathcal{P}_i|}), \mathbf{1}\in\mathbb{R}^{|\mathcal{P}_i|}$$

where $\Delta^{|\mathcal{P}_i|} = \{\mathbf{p}_i^{\text{AAN}} \in \mathbb{R}^{|\mathcal{P}_i|}, \sum_j p_{ij}^{\text{AAN}} = 1, 0 \le p_{ij}^{\text{AAN}} \le 1\}$. Note the expression of $\text{KL}(\mathbf{p}_i^{\text{AAN}}\|\frac{1}{|\mathcal{P}_i|}) = \sum_j p_{ij}\log(|\mathcal{P}_i|p_{ij}) = \sum_j p_{ij}\log(p_{ij}) + \log(|\mathcal{P}_i|)$. There are three constrains to handle, i.e., $\sum_j p_{ij}^{\text{AAN}} = 1, p_{ij}^{\text{AAN}} \ge 0,$ and $p_{ij}^{\text{AAN}} \le 1$. Note that the constraint $p_{ij}^{\text{AAN}} \ge 0$ is enforced by the term $p_{ij}^{\text{AAN}}\log(p_{ij}^{\text{AAN}})$, otherwise the above objective will become infinity. As a result, the constraint $p_{ij}^{\text{AAN}} < 1$ is automatically satisfied due to $\sum_j p_{ij}^{\text{AAN}} = 1$ and $p_{ij}^{\text{AAN}} \ge 0$. Hence, we only need to explicitly tackle the constraint $\sum_j p_{ij}^{\text{AAN}} = 1$. To this end, we define the following Lagrangian

function,

$$\tau L(\mathbf{p}_i^{\text{AAN}}, \mu) = -\sum_{j=1}^{|\mathcal{P}_i|} p_{ij} \mathbf{z}_i(\mathbf{w}_f)^\top \mathbf{z}_j(\mathbf{w}_f) + \tau(\log|\mathcal{P}_i| + \sum_{j=1}^{|\mathcal{P}_i|} p_{ij}^{\text{AAN}} \log(p_{ij}^{\text{AAN}})) + \mu(\sum_j p_{ij}^{\text{AAN}} - 1)$$

where μ is the Lagrangian multiplier for the constraint $\sum_j p_{ij}^{\text{AAN}} = 1$. The optimal solutions satisfy the KKT conditions:

$$\begin{aligned} &- \mathbf{z}_i(\mathbf{w}_f)^\top \mathbf{z}_j(\mathbf{w}_f) + \tau(\log(p_{ij}^{\text{AAN}}(\mathbf{w})) + 1) + \mu = 0, \\ &\sum_j p_{ij}^{\text{AAN}} = 1 \end{aligned}$$

From the first equation, we can derive $p_{ij}^{\text{AAN}} \propto \exp(\frac{\mathbf{z}_i(\mathbf{w}_f)^\top \mathbf{z}_j(\mathbf{w}_f)}{\tau})$. Then according to the second equation, we can conclude that

$$p_{ij}^{\text{AAN}} = \frac{\exp(\frac{\mathbf{z}_i(\mathbf{w}_f)^\top \mathbf{z}_j(\mathbf{w}_f)}{\tau})}{\sum_{j\in\mathcal{P}_i} \exp(\frac{\mathbf{z}_i(\mathbf{w}_f)^\top \mathbf{z}_j(\mathbf{w}_f)}{\tau})} = \frac{\exp(\frac{\mathbf{z}_i^\top(\mathbf{w}_f)\mathbf{z}_j(\mathbf{w}_f)}{\tau})}{\sum\limits_{k\in\mathcal{P}_i} \exp(\frac{\mathbf{z}_i^\top(\mathbf{w}_f)\mathbf{z}_k(\mathbf{w}_f)}{\tau})}$$

Next, we derive the second equivalence in the robust objective, Equation (4).

$$\begin{aligned} \text{RAAN}(\mathbf{w}) :=\ & \frac{1}{C}\sum_{c=1}^{C}\frac{1}{A}\sum_{a=1}^{A}\frac{1}{|\mathcal{D}_a^c|}\sum_{i=1}^{|\mathcal{D}_a^c|}\Bigg(\underbrace{\sum_{j\in\mathcal{P}_i}\frac{\exp(\frac{\mathbf{z}_i^\top(\mathbf{w}_f)\mathbf{z}_j(\mathbf{w}_f)}{\tau})}{\sum\limits_{k\in\mathcal{P}_i}\exp(\frac{\mathbf{z}_i^\top(\mathbf{w}_f)\mathbf{z}_k(\mathbf{w}_f)}{\tau})}}_{p_{ij}^{\text{AAN}}}\ell(\mathbf{w};\mathbf{x}_j,c,a)\Bigg) \\ :=\ & \frac{1}{CA}\sum_{c=1}^{C}\sum_{j\in\mathcal{D}^c}\underbrace{\Bigg(\sum_{i\in\mathcal{P}_j}\frac{1}{|\mathcal{D}_{a_i}^{y_i}|}\frac{\exp(\frac{\mathbf{z}_i^\top(\mathbf{w}_f)\mathbf{z}_j(\mathbf{w}_f)}{\tau})}{\sum\limits_{k\in\mathcal{P}_i}\exp(\frac{\mathbf{z}_i^\top(\mathbf{w}_f)\mathbf{z}_k(\mathbf{w}_f)}{\tau})}\Bigg)}_{p_j^{\text{AAN}}}\ell(\mathbf{w};\mathbf{x}_j,y_j,a_j) \\ :=\ & \frac{1}{CA}\sum_{j\in\mathcal{D}}\underbrace{\Bigg(\sum_{i\in\mathcal{P}_j}\frac{1}{|\mathcal{P}_j|}\frac{\exp(\frac{\mathbf{z}_i^\top(\mathbf{w}_f)\mathbf{z}_j(\mathbf{w}_f)}{\tau})}{\sum\limits_{k\in\mathcal{P}_i}\exp(\frac{\mathbf{z}_i^\top(\mathbf{w}_f)\mathbf{z}_k(\mathbf{w}_f)}{\tau})}\Bigg)}_{p_j^{\text{AAN}}}\ell(\mathbf{w};\mathbf{x}_j,y_j,a_j) \\ \Longleftrightarrow :=\ & \frac{1}{CA}\sum_{j=1}^{|\mathcal{D}|}p_j^{\text{AAN}}\ell(\mathbf{w};\mathbf{x}_j,y_j,a_j) \end{aligned}$$

We finish the derivation. Therefore, RAAN combines the information from the embedding space p_j^{AAN} to promote a more uniform embedding of the classification head.

6.6 Theoretical Analysis

To derive the theoretical analysis, we write the pairwise $\text{RAAN}(\mathbf{w})$, i.e, the first equivalence in Equation (4)

$$\text{RAAN}(\mathbf{w}) := \frac{1}{C}\sum_{c=1}^{C}\frac{1}{A}\sum_{a=1}^{A}\frac{1}{|\mathcal{D}_a^c|}\sum_{i=1}^{|\mathcal{D}_a^c|}\Bigg(\underbrace{\sum_{j\in\mathcal{P}_i}\frac{\exp(\frac{\mathbf{z}_i^\top(\mathbf{w}_f)\mathbf{z}_j(\mathbf{w}_f)}{\tau})}{\sum\limits_{k\in\mathcal{P}_i}\exp(\frac{\mathbf{z}_i^\top(\mathbf{w}_f)\mathbf{z}_k(\mathbf{w}_f)}{\tau})}}_{p_{ij}^{\text{AAN}}}\ell(\mathbf{w};\mathbf{x}_j,c,a)\Bigg)$$

objective as a general compositional form $R(\mathbf{w})$,

$$\mathrm{R}(\mathbf{w}) = \frac{1}{n}\sum_{\mathbf{x}_i \in \mathcal{D}} f(g_{\mathbf{x}_i}(\mathbf{w})) = \mathbb{E}_{\mathbf{x}_i \in \mathcal{D}}[f(g_{\mathbf{x}_i}(\mathbf{w}))]$$

where $f(g) = \frac{g_1}{g_2}$, and $g_{\mathbf{x}_i}(\mathbf{w}) = \frac{n}{AC}\mathbb{E}_{\mathbf{x}_j \in \mathcal{D}}[\exp(\frac{\mathbf{z}_i(\mathbf{w}_f)^\top \mathbf{z}_j(\mathbf{w}_f)}{\tau})\ell_j(\mathbf{w})\mathbb{I}(\mathbf{x}_j \in \mathcal{P}_i), \exp(\frac{\mathbf{z}_i(\mathbf{w}_f)^\top \mathbf{z}_j(\mathbf{w}_f)}{\tau})\mathbb{I}(\mathbf{x}_j \in \mathcal{P}_i)]^\top = \frac{n}{AC}\mathbb{E}_{\mathbf{x}_j \in \mathcal{P}_i}[\exp(\frac{\mathbf{z}_i(\mathbf{w}_f)^\top \mathbf{z}_j(\mathbf{w}_f)}{\tau})\ell_j(\mathbf{w}), \exp(\frac{\mathbf{z}_i(\mathbf{w}_f)^\top \mathbf{z}_j(\mathbf{w}_f)}{\tau})]^\top$, $\forall\, \tau \neq 0$, and $\ell_j(\mathbf{w}) = \ell_j(\mathbf{w}; \mathbf{x}_j, y_j, a_j)$.

Our theoretical analysis follows the same framework as SOAP in (Qi et al., 2021). To make sure the analysis are applicable for both RAAN and RL-RAAN, similar to Equation (7) in section 4, we provide the stochastic gradient estimator for the inner objective for RL-RAAN:

Next, we first introduce the assumptions and provide a lemma to guarantee that $R(\cdot)$ is smooth.

Assumption 1 *Assume that (a) there exists Δ_1 such that $R(\mathbf{w}_1) - \min_{\mathbf{w}} R(\mathbf{w}) \leq \Delta_1$; (b) there exist $M > 0$ such that $\ell(\mathbf{w}; \mathbf{x}_j, y_j, a_j) \leq M$ and $\ell(\mathbf{w}; \mathbf{x}_j, y_j, a_j)$ is C_l-Lipscthiz continuous and L_l-smooth with respect to $\mathbf{w}$ for any $\mathbf{x}_j \in \mathcal{D}$; (c) there exists $V > 0$ such that $\mathbb{E}_{\mathbf{x}_j \sim \mathcal{D}}[\|g(\mathbf{w}; \mathbf{x}_i, \mathbf{x}_j) - g_{\mathbf{x}_i}(\mathbf{w})\|^2] \leq V$, and $\mathbb{E}_{\mathbf{x}_j \sim \mathcal{D}}[\|\nabla g(\mathbf{w}; \mathbf{x}_i, \mathbf{x}_j) - \nabla g_{\mathbf{x}_i}(\mathbf{w})\|^2] \leq V$ for any $\mathbf{x}_i$.*

Lemma 1 *Suppose Assumption 2 holds, $\tau \geq \tau_0$, maxE $= \max\{\exp(1/\tau_0), \exp(-1/\tau_0)\}$, minE $= \min\{\exp(1/\tau_0), \exp(-1/\tau_0)\}$, there exists $u_0 \geq \frac{n \cdot minE}{|\mathcal{P}_i|AC}$, $u_1 = \frac{nM \cdot maxE}{AC}$ and $u_2 = \frac{n \cdot maxE}{AC}$ such that $g_{\mathbf{x}_i}(\mathbf{w}) \in \Omega = \{\mathbf{u} \in \mathbb{R}^2, 0 \leq [\mathbf{u}]_1 \leq u_1, u_0 \leq [\mathbf{u}]_2 \leq u_2\}$, $\forall \mathbf{x}_i \in \mathcal{D}$. In addition, there exists $L > 0$ such that $R(\cdot)$ is L-smooth.*

We first prove the first part $g_i(\mathbf{w}) \in \Omega$. Due to the definition of $g_{\mathbf{x}_i}(\mathbf{w}) = \frac{n}{AC}\mathbb{E}_{\mathbf{x}_j \in \mathcal{D}}[\exp(\frac{\mathbf{z}_i(\mathbf{w}_f)^\top \mathbf{z}_j(\mathbf{w}_f)}{\tau})\ell_j(\mathbf{w})\mathbb{I}(\mathbf{x}_j \in \mathcal{P}_i), \exp(\frac{\mathbf{z}_i(\mathbf{w}_f)^\top \mathbf{z}_j(\mathbf{w}_f)}{\tau})\mathbb{I}(\mathbf{x}_j \in \mathcal{P}_i)]^\top$. As $\mathbf{z}_i(\mathbf{w}_f) = \frac{F(\mathbf{w}_f, \mathbf{x}_i)}{\|F(\mathbf{w}_f, \mathbf{x}_i)\|}$, $-1 \leq \mathbf{z}_i(\mathbf{w}_f)^\top \mathbf{z}_j(\mathbf{w}_f) \leq 1$, $\min\{(\exp(\frac{-1}{\tau}), \exp(\frac{1}{\tau})\} \leq \exp(\frac{\mathbf{z}_i(\mathbf{w}_f)^\top \mathbf{z}_j(\mathbf{w}_f)}{\tau}) \leq \max\{(\exp(\frac{-1}{\tau}), \exp(\frac{1}{\tau})\}$ Therefore, $0 \leq [g_{\mathbf{x}_i}(\mathbf{w})]_1 \leq \frac{n \max\{\exp(1/\tau_0), \exp(-1/\tau_0)\}M}{AC}$ and $\frac{n \min\{\exp(1/\tau_0), \exp(-1/\tau_0)\}}{|\mathcal{P}_i|AC} \leq [g_{\mathbf{x}_i}(\mathbf{w})]_2 \leq \frac{n \max\{\exp(1/\tau_0), \exp(-1/\tau_0)\}}{AC}$ $\forall i, j$. To this end, we need to use the following Lemma 2 and the proof will be presented.

Assumption 2 *Assume that (a) there exists Δ_1 such that $R(\mathbf{w}_1) - \min_{\mathbf{w}} R(\mathbf{w}) \leq \Delta_1$; (b) there exist $M > 0$ such that $\ell(\mathbf{w}; \mathbf{x}_j, y_j, a_j) \leq M$ and $\ell(\mathbf{w}; \mathbf{x}_j, y_j, a_j)$ is C_l-Lipscthiz continuous and L_l-smooth with respect to $\mathbf{w}$ for any $\mathbf{x}_j \in \mathcal{D}$; (c) there exists $V > 0$ such that $\mathbb{E}_{\mathbf{x}_j \sim \mathcal{D}}[\|g(\mathbf{w}; \mathbf{x}_i, \mathbf{x}_j) - g_{\mathbf{x}_i}(\mathbf{w})\|^2] \leq V$, and $\mathbb{E}_{\mathbf{x}_j \sim \mathcal{D}}[\|\nabla g(\mathbf{w}; \mathbf{x}_i, \mathbf{x}_j) - \nabla g_{\mathbf{x}_i}(\mathbf{w})\|^2] \leq V$ for any $\mathbf{x}_i$.*

Lemma 2 *Let $L_f = \frac{4(u_0+u_1)}{u_0^3}, C_f = \frac{u_0+u_1}{u_0^2}, L_g = \frac{10nmaxE}{AC}(C_l + L_l) =, C_g = \frac{nmaxE}{AC}(C_\ell + 2M)$, then $f(\mathbf{u})$ is a L_f-smooth, C_f-Lipschit continuous function for any $\mathbf{u} \in \Omega$, and $\forall i \in [1, \cdots, n]$, $g_{\mathbf{x}_i}$ is a L_g-smooth, C_g-Lipschitz continuous function.*

$$f(\mathbf{u}) = \frac{[\mathbf{u}]_1}{[\mathbf{u}]_2} \quad , \nabla_{\mathbf{u}} f(\mathbf{u}) = \left(\frac{1}{[\mathbf{u}]_2}, -\frac{[\mathbf{u}]_1}{([\mathbf{u}]_2)^2}\right)^\top, \quad \nabla^2_{\mathbf{u}} f(\mathbf{u}) = \begin{pmatrix} 0, -\frac{1}{([\mathbf{u}]_2)^2} \\ -\frac{1}{([\mathbf{u}]_2)^2}, \frac{2[\mathbf{u}]_1}{([\mathbf{u}]_2)^3} \end{pmatrix} \tag{10}$$

Due to the assumption that $\ell(\mathbf{w};\mathbf{x}_i)$ is a L_l-smooth, C_l-Lipschitz continuous function, and $\|\mathbf{z}_i(\mathbf{w}_f)\|^2 = 1, -1 \leq \mathbf{z}_i^\top(\mathbf{w}_f)\mathbf{z}_(\mathbf{w}_f) \leq 1$, we have

$$
\begin{aligned}
\|\nabla^2_{\mathbf{w}} g_i(\mathbf{w})\| &= \|\frac{n}{AC|\mathcal{P}_i|}\sum_{j=1}^{|\mathcal{P}_i|}[\exp(\frac{\mathbf{z}_i(\mathbf{w}_f)^\top \mathbf{z}_j(\mathbf{w}_f)}{\tau})\nabla^2_{\mathbf{w}}\ell_j(\mathbf{w};\mathbf{x}_j,c_j,a_j) \\
&+ 2(\mathbf{z}_i(\mathbf{w}_f)+\mathbf{z}_j(\mathbf{w}_f))\exp(\frac{\mathbf{z}_i(\mathbf{w}_f)^\top \mathbf{z}_j(\mathbf{w}_f)}{\tau})\nabla_{\mathbf{w}}\ell_j(\mathbf{w};\mathbf{x}_j,c_j,a_j) \\
&+ ((\mathbf{z}_i(\mathbf{w}_f)+\mathbf{z}_j(\mathbf{w}_f))^2+2)\exp(\frac{\mathbf{z}_i(\mathbf{w}_f)^\top \mathbf{z}_j(\mathbf{w}_f)}{\tau})\ell_j(\mathbf{w};\mathbf{x}_j,c_j,a_j)\|] \\
&\overset{(a)}{\leq} \frac{n}{AC}\frac{1}{|\mathcal{P}_i|}\sum_{j=1}^{|\mathcal{P}_i|}\|[\exp(\frac{\mathbf{z}_i(\mathbf{w}_f)^\top \mathbf{z}_j(\mathbf{w}_f)}{\tau})\nabla^2_{\mathbf{w}}\ell_j(\mathbf{w};\mathbf{x}_j,c_j,a_j) \\
&+ 2(\mathbf{z}_i(\mathbf{w}_f)+\mathbf{z}_j(\mathbf{w}_f))\exp(\frac{\mathbf{z}_i(\mathbf{w}_f)^\top \mathbf{z}_j(\mathbf{w}_f)}{\tau})\nabla_{\mathbf{w}}\ell_j(\mathbf{w};\mathbf{x}_j,c_j,a_j) \\
&+ ((\mathbf{z}_i(\mathbf{w}_f)+\mathbf{z}_j(\mathbf{w}_f))^2+2)\exp(\frac{\mathbf{z}_i(\mathbf{w}_f)^\top \mathbf{z}_j(\mathbf{w}_f)}{\tau})\ell_j(\mathbf{w};\mathbf{x}_j,c_j,a_j)\| \\
&\leq \frac{n\text{maxE}}{AC}(L_l+10C_l) \leq \frac{10n\text{maxE}}{AC}(C_l+L_l) = L_g
\end{aligned}
\tag{11}
$$

where (a) applies the convexity of $\|\cdot\|$ and $\|a+b\| \leq \|a\|+\|b\|$. Similarly, the following equations hold in terms of the continuous of inner objective $g_{\mathbf{x}_i}$,

$$
\begin{aligned}
\|\nabla_{\mathbf{w}} g_{\mathbf{x}_i}(\mathbf{w})\| &= \|\frac{n}{AC|\mathcal{P}_i|}\sum_{j=1}^{|\mathcal{P}_i|}[\exp(\frac{\mathbf{z}_i(\mathbf{w}_f)^\top \mathbf{z}_j(\mathbf{w}_f)}{\tau})\nabla_{\mathbf{w}}\ell_j(\mathbf{w};\mathbf{x}_j,c_j,a_j) \\
&\quad + (\mathbf{z}_i(\mathbf{w}_f)+\mathbf{z}_j(\mathbf{w}_f))\exp(\frac{\mathbf{z}_i(\mathbf{w}_f)^\top \mathbf{z}_j(\mathbf{w}_f)}{\tau})\ell_j(\mathbf{w};\mathbf{x}_j,c_j,a_j)\|] \\
&\leq \frac{n}{AC}(\text{maxE}C_\ell + 2\text{maxE}M)) = \frac{n\text{maxE}}{AC}(C_\ell+2M) = C_g
\end{aligned}
\tag{12}
$$

$$
\begin{aligned}
\|\nabla f(\mathbf{u})\| &\leq \sqrt{\frac{1}{[\mathbf{u}]_2^2}+\frac{[\mathbf{u}]_1^2}{[\mathbf{u}]_2^4}} \leq \frac{u_0+u_1}{u_0^2} = C_f \\
\|\nabla^2 f(\mathbf{u})\| &\leq \sqrt{\frac{2}{[\mathbf{u}]_2^4}+4\frac{[\mathbf{u}]_1^2}{[\mathbf{u}]_2^6}} \leq \frac{4(u_0+u_1)}{u_0^3} = L_f
\end{aligned}
\tag{13}
$$

Since $P(\mathbf{w}) = \frac{1}{n}\sum_{\mathbf{x}_i\in\mathcal{D}} f(g_i(\mathbf{w}))$. We first show $R_i(\mathbf{w}) = f(g_i(\mathbf{w}))$ is smooth. To see this,

$$
\begin{aligned}
&\|\nabla R_i(\mathbf{w}) - \nabla R_i(\mathbf{w}')\| = \|\nabla g_i(\mathbf{w})^\top \nabla f(g_i(\mathbf{w})) - \nabla g_i(\mathbf{w}')^\top \nabla f(g_i(\mathbf{w}'))\| \\
&\leq \|\nabla g_i(\mathbf{w})^\top \nabla f(g_i(\mathbf{w})) - \nabla g_i(\mathbf{w}')^\top \nabla f(g_i(\mathbf{w}))\| \\
&+ \|\nabla g_i(\mathbf{w}')^\top \nabla f(g_i(\mathbf{w})) - \nabla g_i(\mathbf{w}')^\top \nabla f(g_i(\mathbf{w}'))\| \\
&\leq C_f L_g\|\mathbf{w}-\mathbf{w}'\| + C_g L_f C_g\|\mathbf{w}-\mathbf{w}'\| = (C_f L_g + L_f C_g^2)\|\mathbf{w}-\mathbf{w}'\|.
\end{aligned}
$$

Hence $R(\mathbf{w})$ is also $L = (C_f L_g + L_f C_g^2)$-smooth.

6.7 Proof of Theorem 1 (SCRAAN with SGD-Style Update)

Lemma 3 *With $\alpha \leq 1/2$, running T iterations of SCRAAN (SGD-style) updates, we have*

$$
\frac{\alpha}{2}\mathbb{E}[\sum_{t=1}^{T}\|\nabla R(\mathbf{w}_t)\|^2] \leq \mathbb{E}[\sum_t (R(\mathbf{w}_t) - R(\mathbf{w}_{t+1}))] + \frac{\alpha C_1}{2}\mathbb{E}[\sum_{t=1}^{T}\|g_{i_t}(\mathbf{w}_t) - \mathbf{u}_{i_t}\|^2] + \alpha^2 T C_2,
$$

where i_t denotes the index of the sampled positive data at iteration t, C_1 and C_2 are proper constants.

Our key contribution is the following lemma that bounds the second term in the above upper bound.

Lemma 4 *Suppose Assumption 2 holds, with* $\mathbf{u}$ *initialized inner objective stochastic estimator for every* $\mathbf{x}_i \in \mathcal{D}$ *we have*

$$\mathbb{E}[\sum_{t=1}^{T} \|g_{i_t}(\mathbf{w}_t) - \mathbf{u}_{i_t}\|^2] \leq \frac{nV}{\gamma} + \gamma VT + 2\frac{n^2\alpha^2 TC_3}{\gamma^2}, \tag{14}$$

where C_3 *is a proper constant.*

Remark: The innovation of proving the above lemma is by grouping $\mathbf{u}_{i_t}, t = 1, \ldots, T$ into n groups corresponding to the n samples AAN, and then establishing the recursion of the error $\|g_{i_t}(\mathbf{w}_t) - \mathbf{u}_{i_t}\|^2$ within each group, and then summing up these recursions together.

6.7.1 Proof of Lemma 3

[Proof of Lemma 3] To make the proof clear, we write $\nabla g_{i_t}(\mathbf{w}; \xi) = \nabla g(\mathbf{w}_t; \mathbf{x}_{i_t}, \xi), \xi \sim \mathcal{P}_{i_t}$. Let $\mathbf{u}_{i_t}$ denote the updated $\mathbf{u}$ vector at the t-th iteration for the selected positive data i_t.

$$\begin{aligned}
&R(\mathbf{w}_{t+1}) - R(\mathbf{w}_t) \leq \nabla R(\mathbf{w}_t)^\top(\mathbf{w}_{t+1} - \mathbf{w}_t) + \frac{L}{2}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2 \\
&= -\alpha\|\nabla R(\mathbf{w}_t)\|^2 + \alpha\nabla R(\mathbf{w}_t)^\top(\nabla R(\mathbf{w}_t) - \nabla g_{i_t}^\top(\mathbf{w}_t; \xi)\nabla f(\mathbf{u}_{i_t})) + \frac{\alpha^2\|G(\mathbf{w}_t)\|^2 L}{2} \\
&\leq -\alpha\|\nabla R(\mathbf{w}_t)\|^2 + \alpha\nabla R(\mathbf{w}_t)^\top(\nabla R(\mathbf{w}_t) - \nabla g_{i_t}^\top(\mathbf{w}_t; \xi)\nabla f(\mathbf{u}_{i_t})) + \alpha^2 C_2
\end{aligned}$$

where $C_2 = \|G(\mathbf{w}_t)\|^2 L/2 \leq C_g^2 C_f^2 L/2$.

Taking expectation on both sides, we have

$$\begin{aligned}
\mathbb{E}_t[R(\mathbf{w}_{t+1})] &\leq \mathbb{E}_t[R(\mathbf{w}_t) + \nabla R(\mathbf{w}_t)^\top(\mathbf{w}_{t+1} - \mathbf{w}_t) + \frac{L}{2}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2] \\
&= \mathbb{E}_t[R(\mathbf{w}_t) - \alpha\|\nabla R(\mathbf{w}_t)\|^2 + \alpha\nabla R(\mathbf{w}_t)^\top(\nabla R(\mathbf{w}_t) - \nabla g_{i_t}(\mathbf{w}_t; \xi)^\top\nabla f(\mathbf{u}_{i_t}))] + \alpha^2 C_2 \\
&= R(\mathbf{w}_t) - \alpha\|\nabla R(\mathbf{w}_t)\|^2 + \alpha\nabla R(\mathbf{w}_t)^\top(\mathbb{E}_t[\nabla R(\mathbf{w}_t) - \nabla g_{i_t}(\mathbf{w}_t; \xi)^\top\nabla f(\mathbf{u}_{i_t})]) + \alpha^2 C_2
\end{aligned}$$

where $\mathbb{E}_t$ means taking expectation over i_t, ξ given $\mathbf{w}_t$.
Noting that $\nabla R(\mathbf{w}_t) = \mathbb{E}_{i_t,\xi}[\nabla g_{i_t}(\mathbf{w}_t; \xi)^\top \nabla f(g_{i_t}(\mathbf{w}_t))]$, where i_t and ξ are independent.

$$\begin{aligned}
&\mathbb{E}_t[R(\mathbf{w}_{t+1})] - R(\mathbf{w}_t) \\
&\leq -\alpha\|\nabla R(\mathbf{w}_t)\|^2 + \alpha\nabla R(\mathbf{w}_t)^\top(\mathbb{E}_t[\nabla g_{i_t}(\mathbf{w}_t; \xi)^\top\nabla f(g_{i_t}(\mathbf{w}_t))] - \mathbb{E}_t[\nabla g_{i_t}(\mathbf{w}_t; \xi)^\top\nabla f(\mathbf{u}_{i_t})]) + \alpha^2 C_2 \\
&= -\alpha\|\nabla R(\mathbf{w}_t)\|^2 + \mathbb{E}_t[\alpha\nabla R(\mathbf{w}_t)^\top(\nabla g_{i_t}(\mathbf{w}_t; \xi)^\top\nabla f(g_{i_t}(\mathbf{w}_t)) - \nabla g_{i_t}(\mathbf{w}_t; \xi)^\top\nabla f(\mathbf{u}_{i_t}))] + \alpha^2 C_2 \\
&\overset{(a)}{\leq} -\alpha\|\nabla R(\mathbf{w}_t)\|^2 + \mathbb{E}_t[\frac{\alpha}{2}\|\nabla R(\mathbf{w}_t)\|^2 + \frac{\alpha}{2}\|\nabla g_{i_t}(\mathbf{w}_t; \xi)^\top\nabla f(g_{i_t}(\mathbf{w}_t)) - \nabla g_{i_t}(\mathbf{w}_t; \xi)^\top\nabla f(\mathbf{u}_{i_t}))\|^2 + \alpha^2 C_2 \\
&\overset{(b)}{\leq} -\alpha\|\nabla R(\mathbf{w}_t)\|^2 + \mathbb{E}_t[\frac{\alpha}{2}\|\nabla R(\mathbf{w}_t)\|^2 + \frac{\alpha C_1}{2}\|g_{i_t}(\mathbf{w}_t) - \mathbf{u}_{i_t}\|^2 + \alpha^2 C_2 \\
&= -(\alpha - \frac{\alpha}{2})\|\nabla R(\mathbf{w}_t)\|^2 + \frac{\alpha C_1}{2}\mathbb{E}_t[\|g_{i_t}(\mathbf{w}_t) - \mathbf{u}_{i_t}\|^2] + \alpha^2 C_2
\end{aligned}$$

where the equality (a) is due to $ab \leq a^2/2 + b^2/2$ and the inequality (b) uses the factor $\|\nabla g_{i_t}(\mathbf{w}_t; \xi)\| \leq C_l$ and ∇f is L_f-Lipschitz continuous for $\mathbf{u}, g_i(\mathbf{w}) \in \Omega$ and $C_1 = C_l^2 C_f^2$. Hence we have,

$$\frac{\alpha}{2}\|\nabla R(\mathbf{w}_t)\|^2 \leq R(\mathbf{w}_t) - \mathbb{E}_t[R(\mathbf{w}_{t+1})] + \frac{\alpha C_1}{2}\mathbb{E}_t[\|g_{i_t}(\mathbf{w}_t) - \mathbf{u}_{i_t}\|^2] + \alpha^2 C_2$$

Taking summation and expectation over all randomness, we have

$$\frac{\alpha}{2}\mathbb{E}[\sum_{t=1}^{T}\|\nabla R(\mathbf{w}_t)\|^2] \leq \mathbb{E}[\sum_t (R(\mathbf{w}_t) - R(\mathbf{w}_{t+1}))] + \frac{\alpha C_1}{2}\mathbb{E}[\sum_{t=1}^{T}\|g_{i_t}(\mathbf{w}_t) - \mathbf{u}_{i_t}\|^2] + \alpha^2 C_2 T$$

6.7.2 PROOF OF LEMMA 4

Let i_t denote the selected data i_t at t-th iteration. We will divide $\{1, \ldots, T\}$ into n groups with the i-th group given by $\mathcal{T}_i = \{t_1^i, \ldots, t_k^i \ldots, \}$, where t_k^i denotes the iteration that the i-th index data is selected at the k-th time for updating $\mathbf{u}$. Let us define $\phi(t) : [T] \to [n] \times [T]$ that maps the selected data into its group index and within group index, i.e, there is an one-to-one correspondence between index t and selected data i and its index within $\mathcal{T}_i$. Below, we use notations a_i^k to denote $a_{t_k^i}$. Let $T_i = |\mathcal{T}_i|$. Hence, $\sum_{i=1}^{n_+} T_i = T$.

[Proof of Lemma 4] To prove Lemma 4, we first introduce another lemma that establishes a recursion for $\|\mathbf{u}_{i_t} - g_{i_t}(\mathbf{w}_t)\|^2$, whose proof is presented later.

Lemma 5 *By the updates of SCRAAN Adam-style or SGD-style with a sample $\mathbf{x}_i \in \mathcal{D}$, and, $\xi \in \mathcal{P}_i$, the following equation holds for $\forall\ t \in 1, \cdots, T$*

$$\begin{aligned}\mathbb{E}_t[\|\mathbf{u}_{i_t} - g_{i_t}(\mathbf{w}_t)\|^2] &\overset{\phi(t)}{=} \mathbb{E}_t[\|\mathbf{u}_i^k - g_i(\mathbf{w}_i^k)\|^2] \\ &\leq (1-\gamma)\|\mathbf{u}_i^{k-1} - g_i(\mathbf{w}_i^{k-1})\|^2 + \gamma^2 V + \gamma^{-1}\alpha^2 n^2 C_3\end{aligned} \tag{15}$$

where $\mathbb{E}_t$ denotes the conditional expectation conditioned on history before t_{k-1}^i.

Then, by mapping every i_t to its own group and make use of Lemma 5, we have

$$\mathbb{E}[\sum_{k=0}^{K_i} \|\mathbf{u}_i^k - g_i^k(\mathbf{w}_i^k)\|^2] \leq \mathbb{E}\left[\frac{[\|\mathbf{u}_i^0 - g_i(\mathbf{w}_i^0)\|^2]}{\gamma} + \gamma V T_i + \gamma^{-2} n^2 C_3 \alpha^2 T_i\right] \tag{16}$$

where $\mathbf{u}_i^0$ is the initial vector for $\mathbf{u}_i$, which can be computed by a mini-batch averaging estimator of $g_i(\mathbf{w}_0)$. Thus

$$\begin{aligned}\mathbb{E}[\sum_{t=1}^{T} \|g_{i_t}(\mathbf{w}_t) - \mathbf{u}_{i_t}\|^2] &\overset{\phi(t)}{=} \mathbb{E}[\sum_{i=1}^{n}\sum_{k=0}^{K_i} \|\mathbf{u}_i^k - g_i^k(\mathbf{w}_i^k)\|^2] \\ &\leq \sum_{i=1}^{n}\left\{\frac{[\|\mathbf{u}_i^0 - g_i^0(\mathbf{w}_i^0)\|^2]}{\gamma} + \gamma V \mathbb{E}[T_i] + \gamma^{-2} n^2 C_3 \alpha^2 \mathbb{E}[T_i]\right\} \\ &\leq \frac{nV}{\gamma} + \gamma V T + \frac{n^2 \alpha^2 T C_3}{\gamma^2}\end{aligned}$$

6.7.3 PROOF OF LEMMA 5

We first introduce the following lemma, whose proof is presented later.

Lemma 6 *Suppose the sequence generated in the training process using the positive sample i is $\{\mathbf{w}_{i_1}^i, \mathbf{w}_{i_2}^i, ..., \mathbf{w}_{i_{T_i}}^i\}$, where $0 < i_1 < i_2 < \cdots < i_{T_i} \leq T$, then $\mathbb{E}_{|i_k}[i_{k+1} - i_k] \leq n_+$, and, $\mathbb{E}_{|i_k}[(i_{k+1} - i_k)^2] \leq 2n^2, \forall k$.*

Define $\widetilde{g}_{i_t}(\mathbf{w}_t) = g(\mathbf{w}_t, \mathbf{x}_{i_t}, \xi)$. Let $\prod_\Omega(\cdot) : \mathbb{R}^2 \to \Omega$ denotes the projection operator. By the updates of $\mathbf{u}_{i_t}$, we have $\mathbf{u}_{i_t} = \mathbf{u}_i^k = \prod_\Omega[(1-\gamma)\mathbf{u}_i^{k-1} + \gamma \tilde{g}_{i_t}(\mathbf{w}_t)]$.

$$
\begin{aligned}
&\mathbb{E}_t[\|\mathbf{u}_{i_t} - g_{i_t}(\mathbf{w}_t)\|^2] \overset{\phi(t)}{=} \mathbb{E}[\|\mathbf{u}_i^k - g_i(\mathbf{w}_i^k)\|^2] \\
&= \mathbb{E}_t[\|\prod_{\Omega}((1-\gamma)\mathbf{u}_i^{k-1} + \gamma\widetilde{g}_i(\mathbf{w}_i^k)) - \prod_{\Omega}(g_i(\mathbf{w}_t))\|^2] \\
&\leq \mathbb{E}_t[\|((1-\gamma)\mathbf{u}_i^{k-1} + \gamma\widetilde{g}_i(\mathbf{w}_i^k) - g_i(\mathbf{w}_t)\|^2] \\
&\leq \mathbb{E}_t[\|((1-\gamma)(\mathbf{u}_i^{k-1} - g_i(\mathbf{w}_i^{k-1})) + \gamma(\widetilde{g}_i(\mathbf{w}_i^k) - g_i(\mathbf{w}_i^k)) + (1-\gamma)(g_i(\mathbf{w}_i^{k-1}) - g_i(\mathbf{w}_i^k))\|^2] \\
&\leq \mathbb{E}_t[\|((1-\gamma)(\mathbf{u}_i^{k-1} - g_i(\mathbf{w}_i^{k-1})) + (1-\gamma)(g_i(\mathbf{w}_i^{k-1}) - g_i(\mathbf{w}_i^k))\|^2] + \gamma^2 V \\
&\leq [(1-\gamma)^2(1+\gamma)\|\mathbf{u}_i^{k-1} - g_i(\mathbf{w}_i^{k-1})\|^2] + \gamma^2 V + \frac{(1+\gamma)(1-\gamma)^2}{\gamma} C_g \mathbb{E}[\|\mathbf{w}_i^k - \mathbf{w}_i^{k-1}\|^2] \\
&\leq [(1-\gamma)\|\mathbf{u}_i^{k-1} - g_i(\mathbf{w}_i^{k-1})\|^2] + \gamma^2 V + \gamma^{-1}\alpha^2 C_g \mathbb{E}_t[\|\sum_{t=t_{k-1}^i}^{t_k^i - 1} \nabla g_{i_t}(\mathbf{w}_t;\xi)\nabla f(\mathbf{u}_{i_t})\|^2] \\
&\leq [(1-\gamma)\|\mathbf{u}_i^{k-1} - g_i(\mathbf{w}_i^{k-1})\|^2] + \gamma^2 V + \gamma^{-1}\alpha^2 C_g \mathbb{E}_t[(t_k^i - t_{k-1}^i)^2] C_g^2 C_f^2)] \\
&\overset{(a)}{\leq} \mathbb{E}[(1-\gamma)\|\mathbf{u}_i^{k-1} - g_i(\mathbf{w}_i^{k-1})\|^2] + \gamma^2 V + 2\gamma^{-1}\alpha^2 n^2 C_g^3 C_f^2 \\
&\leq [(1-\gamma)\|\mathbf{u}_i^{k-1} - g_i(\mathbf{w}_i^{k-1})\|^2] + \gamma^2 V + \gamma^{-1}\alpha^2 n^2 C_3
\end{aligned}
$$

where the inequality (a) is due to that $t_k^i - t_{k-1}^i$ is a geometric distribution random variable with $p = 1/n$, i.e., $\mathbb{E}_{|t_{k-1}^i}[(t_k^i - t_{k-1}^i)^2] \leq 2/p^2 = 2n^2$, by Lemma 6. The last equality hold by defining $C_3 = 2C_g^3 C_f^2$.

6.7.4 Proof of Lemma 6

Proof of Lemma 6. Denote the random variable $\Delta_k = i_{k+1} - i_k$ that represents the iterations that the ith positive sample has been randomly selected for the $k+1$-th time conditioned on i_k. Then Δ_k follows a Geometric distribution such that $\Pr(\Delta_k = j) = (1-p)^{j-1}p$, where $p = \frac{1}{n}$, $j = 1, 2, 3, \cdots$. As a result, $\mathbb{E}[\Delta_k | i_k] = 1/p = n$. $\mathbb{E}[\Delta_k^2 | i_k] = \text{Var}(\Delta_k) + \mathbb{E}[\Delta_k | i_k]^2 = \frac{1-p}{p^2} + \frac{1}{p^2} \leq \frac{2}{p^2} = 2n^2$.

6.8 Proof of Theorem 1 (SCRAAN with Adam-Style Update)

We first provide two useful lemmas, whose proof are presented later.

Lemma 7 *Assume assumption 2 holds*

$$
\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2 \leq \alpha^2 d(1-\eta_2)^{-1}(1-\tau)^{-1} \tag{17}
$$

where d is the dimension of $\mathbf{w}$, $\eta_1 < \sqrt{\eta_2} < 1$, and $\tau := \eta_1^2/\eta_2$.

Lemma 8 *With $c = (1 + (1-\eta_1)^{-1})\epsilon^{-\frac{1}{2}} C_g^2 L_f^2$, running T iterations of SOAP (Adam-style) updates, we have*

$$
\begin{aligned}
&\sum_{t=1}^{T} \frac{\alpha(1-\eta_1)(\epsilon + C_g^2 C_f^2)^{-1/2}}{2} \|\nabla R(\mathbf{w}_t)\|^2 \leq \mathbb{E}[\mathbb{V}_1] - \mathbb{E}[\mathbb{V}_{T+1}] \\
&+ 2\eta_1 L\alpha^2 T d(1-\eta_1)^{-1}(1-\eta_2)^{-1}(1-\tau)^{-1} + L\alpha^2 T d(1-\eta_2)^{-1}(1-\tau)^{-1} \\
&+ 2(1-\eta_1)^{-1}\alpha C_g^2 C_f^2 \sum_{i'=1}^{d}((\epsilon + \hat{v}_0^{i'})^{-1/2}) + c\alpha \sum_{t=1}^{T} \mathbb{E}_t[\|g_{i_t}(\mathbf{w}_t) - \mathbf{u}_{i_t}\|^2]
\end{aligned} \tag{18}
$$

where $\mathbb{V}_{t+1} = P(\mathbf{w}_{t+1}) - c_{t+1}\langle \nabla P(\mathbf{w}_t), D_{t+1} h_{t+1} \rangle$.

According to Lemma 8 and plugging Lemma 4 into equation (18), we have

$$\begin{aligned}
&\sum_{t=1}^{T} \frac{\alpha(1-\eta_1)(\epsilon + C_g^2 C_f^2)^{-1/2}}{2} \|\nabla R(\mathbf{w}_t)\|^2 \\
&\leq \mathbb{E}[\mathbb{V}_1] - \mathbb{E}[\mathbb{V}_{T+1}] + 2\eta_1 L\alpha^2 T d(1-\eta_1)^{-1}(1-\eta_2)^{-1}(1-\tau)^{-1} + L\alpha^2 dT(1-\eta_2)^{-1}(1-\tau)^{-1} \\
&+ 2c\alpha C_g^2 C_f^2 \sum_{i'=1}^{d} (\epsilon + \hat{v}_0^{i'})^{-1/2} + c\alpha(\frac{nV}{\gamma} + 2\gamma VT + \frac{2C_g n^2 C_3 \alpha^2 T}{\gamma^2})
\end{aligned} \tag{19}$$

Let $\eta' = (1-\eta_2)^{-1}(1-\tau)^{-1}, \eta^{''} = (1-\eta_1)^{-1}(1-\eta_2)^{-1}(1-\tau)^{-1}$, and $\widetilde{\eta} = (1-\eta_1)^{-2}(1-\eta_2)^{-1}(1-\tau)^{-1}$. As $(1-\eta_1)^{-1} \geq 1, (1-\eta_2)^{-1} \geq 1$, then $\widetilde{\eta} \geq \eta^{''} \geq \eta' \geq 1$.

Then by rearranging terms in Equation (19), dividing $\alpha T(1+\eta_1)(\epsilon + C_g^2 C_f^2)^{-1/2}$ on both sides and suppress constants, $C_g, L_g, C_3, L, C_f, L_f, V, \epsilon$ into big O, we get

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} \|\nabla R(\mathbf{w}_t)\|^2 &\leq \frac{1}{\alpha T(1-\eta_1)} O\Big(\mathbb{E}[\mathbb{V}_1] - \mathbb{E}[\mathbb{V}_{T+1}] + \eta^{''}\eta_1\alpha^2 Td + \eta^{'}\alpha^2 Td + \alpha \sum_{i'=1}^{d} (\epsilon + \hat{v}_0^{i'})^{-1/2} \\
&\quad + \frac{c\alpha n}{\gamma} + c\alpha\gamma T + \frac{c\alpha^3 n^2 T}{\gamma^2}\Big) \\
&\overset{(a)}{\leq} \frac{1}{\alpha T(1-\eta_1)} O\Big(\mathbb{E}[\mathbb{V}_1] - \mathbb{E}[\mathbb{V}_{T+1}] + \eta^{''}\eta_1\alpha^2 Td + \eta^{'}\alpha^2 Td + \alpha d(\epsilon + C_f C_g)^{-1/2} \\
&\quad + \frac{c\alpha n}{\gamma} + c\alpha\gamma T + \frac{c\alpha^3 n^2 T}{\gamma^2}\Big) \\
&\overset{(b)}{\leq} \frac{\widetilde{\eta}}{\alpha T} O\Big(\mathbb{E}[\mathbb{V}_1] - [\mathbb{V}_{T+1}] + (1+\eta_1)\alpha^2 Td + \alpha d + \frac{c\alpha n}{\gamma} + c\alpha\gamma T + \frac{c\alpha^3 n^2 T}{\gamma^2}\Big)
\end{aligned} \tag{20}$$

where the inequality (a) is due to $\hat{v}_0^{i'} = G^{i'}(\mathbf{w}_0)^2 \leq \|G(\mathbf{w}_0)\|^2 \leq C_f^2 C_g^2$. The last inequality (b) is due to $\widetilde{\eta} \geq \eta^{''} \geq \eta' \geq 1$.

Moreover, by the definition of $\mathbb{V}$ and $\mathbf{w}_0 = \mathbf{w}_1$, we have

$$\begin{aligned}
\mathbb{E}[\mathbb{V}_1] &= R(\mathbf{w}_1) - c_1\langle \nabla R(\mathbf{w}_0), D_1 h_1\rangle \leq R(\mathbf{w}_1) + c_1\|\nabla R(\mathbf{w}_0)\|\|\mathbf{w}_1 - \mathbf{w}_0\|\frac{1}{\alpha} = R(\mathbf{w}_1) \\
-\mathbb{E}[\mathbb{V}_{T+1}] &\leq -R(\mathbf{w}_{T+1}) + c_{T+1}\langle \nabla R(\mathbf{w}_T), D_T h_T\rangle \\
&\leq -\min_{\mathbf{w}} R(\mathbf{w}) + c_{T+1}\|\nabla R(\mathbf{w}_{t-1})\|\|\mathbf{w}_{t+1} - \mathbf{w}_t\|\frac{1}{\alpha} \\
&\overset{(a)}{\leq} -\min_{\mathbf{w}} R(\mathbf{w}) + (1-\eta_1)^{-1}\alpha\sqrt{d}(1-\eta_2)^{-1/2}(1-\tau)^{-1/2} \\
&\overset{(b)}{\leq} -\min_{\mathbf{w}} R(\mathbf{w}) + \widetilde{\eta}\sqrt{d}\alpha
\end{aligned} \tag{21}$$

where the inequality (a) is due to Lemma 7 and $c_{T+1} \leq (1-\eta_1)^{-1}\alpha$ in equation (34). The inequality (b) is due to $(1-\eta_1)^{-1}(1-\eta_2)^{-1/2}(1-\tau)^{-1/2} \leq (1-\eta_1)^{-1}(1-\eta_2)^{-1}(1-\tau)^{-1} \leq \eta^{''} \leq \widetilde{\eta}$. Thus $\mathbb{E}[\mathbb{V}_1] - \mathbb{E}[\mathbb{V}_{T+1}] \leq P(\mathbf{w}_1) - \min_{\mathbf{w}} P(\mathbf{w}) + \widetilde{\eta}\sqrt{d}\alpha \leq \Delta_1 + \widetilde{\eta}\sqrt{d}\alpha$ by combining equation (20) and (21).
Then we have

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} \|\nabla R(\mathbf{w}_t)\|^2 &\leq \widetilde{\eta} O\Big(\frac{\Delta_1 + \widetilde{\eta}\sqrt{d}\alpha}{\alpha T} + (1+\eta_1)\alpha d + \frac{d}{T} + \frac{nc}{T\gamma} + c\gamma + \frac{\alpha^2 n^2}{\gamma^2}\Big) \\
&\overset{(a)}{\leq} \widetilde{\eta} O\Big(\frac{\Delta_1 n^{2/5}}{T^{2/5}} + \frac{\widetilde{\eta}\sqrt{d}}{T} + \frac{(1+\eta_1)d}{n^{2/5}T^{3/5}} + \frac{d}{T} + \frac{cn^{3/5}}{T^{3/5}} + 2\frac{cn^{2/5}}{T^{2/5}}\Big) \\
&\overset{(b)}{\leq} O(\frac{n^{2/5}}{T^{2/5}})
\end{aligned} \tag{22}$$

The inequality (a) is due to $\gamma = \frac{n^{2/5}}{T^{2/5}}$, $\alpha = \frac{1}{n^{2/5}T^{3/5}}$. In inequality (b), we further compress the Δ_1, η_1, $\widetilde{\eta}$, c into big O and $\gamma \leq 1 \rightarrow n^{2/5} \leq T^{2/5}$.

6.8.1 PROOF OF LEMMA 7

This proof is following the proof of Lemma 4 in (Chen et al., 2021).

Choosing $\eta_1 < 1$ and defining $\tau = \frac{\eta_1^2}{\eta_2}$, with the Adam-style (Algorithm 3) updates of SOAP that $h_{t+1} = \eta_1 h_t + (1-\eta_1)G(\mathbf{w}_t)$, we can verify for every dimension l,

$$\begin{aligned}
|h_{t+1}^l| &= |\eta_1 h_t^l + (1-\eta_1)G^l(\mathbf{w}_t)| \leq \eta_1 |h_t^l| + |G^l(\mathbf{w}_t)| \\
&\leq \eta_1(\eta_1 |h_{t-1}^l| + |G^l(\mathbf{w}_{t-1})|) + |G^l(\mathbf{w}_t)| \\
&\leq \sum_{p=0}^{t} \eta_1^{t-p} |G^l(\mathbf{w}_p)| = \sum_{p=0}^{t} \sqrt{\tau}^{t-p} \sqrt{\eta_2}^{t-p} |G^l(\mathbf{w}_p)| \\
&\leq \Big(\sum_{p=0}^{t} \tau^{t-p}\Big)^{\frac{1}{2}} \Big(\sum_{p=0}^{t} \eta_2^{t-p} (G^l(\mathbf{w}_p))^2\Big)^{\frac{1}{2}} \\
&\leq (1-\tau)^{-\frac{1}{2}} \Big(\sum_{p=0}^{t} \eta_2^{t-p} (G^l(\mathbf{w}_t))^2\Big)^{\frac{1}{2}}
\end{aligned} \tag{23}$$

where $\mathbf{w}^l$ is the lth dimension of $\mathbf{w}$, the third inequality follows the Cauchy-Schwartz inequality. For the lth dimension of $\hat{v}$, $\hat{v}_t^l$, first we have $\hat{v}_1^l \geq (1-\eta_2)(G^l(\mathbf{w}_1)^2)$. Then since

$$\hat{v}_{t+1}^l \geq \eta_t \hat{v}_t^l + (1-\eta_2)(G^l(\mathbf{w}_t))^2$$

by induction we have

$$\hat{v}_{t+1}^l \geq (1-\eta_2) \sum_{p=0}^{t} \eta_2^{t-p} (G^l(\mathbf{w}_t))^2 \tag{24}$$

Using equation (23) and equation (24), we have

$$\begin{aligned}
|h_{t+1}^l|^2 &\leq (1-\tau)^{-1} \Big(\sum_{p=0}^{t} \eta_2^{t-p} (G^l(\mathbf{w}_t))^2\Big) \\
&\leq (1-\eta_2)^{-1}(1-\tau)^{-1} \hat{v}_{t+1}^l
\end{aligned} \tag{25}$$

Then follow the Adam-style update in Algorithm 3, we have

$$\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2 = \alpha^2 \sum_{l=1}^{d} (\epsilon + \hat{v}_{t+1}^l)^{-1} |h_{t+1}^l|^2 \leq \alpha^2 d (1-\eta_2)^{-1}(1-\tau)^{-1} \tag{26}$$

which completes the proof.

6.8.2 PROOF OF LEMMA 8

To make the proof clear, we make some definitions the same as the proof of Lemma 3. Denote by $\nabla g_{i_t}(\mathbf{w}_t; \xi) = \nabla g(\mathbf{w}_t; \mathbf{x}_{i_t}, \xi), \xi \sim \mathcal{P}_{i_t}$, where i_t is a positive sample randomly generated from $\mathcal{D}$ at t-th iteration, and ξ is a random sample that generated from $\mathcal{D}$ at t-th iteration. It is worth to notice that i_t and ξ are independent. $\mathbf{u}_{i_t}$ denote the updated $\mathbf{u}$ vector at the t-th iteration for the selected positive data i_t.

$$\begin{aligned}
R(\mathbf{w}_{t+1}) &\leq R(\mathbf{w}_t) + \nabla R(\mathbf{w}_t)^\top (\mathbf{w}_{t+1} - \mathbf{w}_t) + \frac{L}{2} \|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2 \\
&\leq R(\mathbf{w}_t) - \alpha \nabla R(\mathbf{w}_t)^\top (D_{t+1} h_{t+1}) + \alpha^2 d (1-\eta_2)^{-1}(1-\tau)^{-1} L/2
\end{aligned}$$

where $D_{t+1} = \frac{1}{\sqrt{\epsilon I + \hat{\mathbf{v}}_{t+1}}}$, $h_{t+1} = \eta_1 h_t + (1-\eta_1)\nabla g_{i_t}^\top(\mathbf{w}_t;\xi)\nabla f(\mathbf{u}_{i_t})$ and the second inequality is due to Lemma 7. Taking expectation on both sides, we have

$$\mathbb{E}_t[R(\mathbf{w}_{t+1})] \leq R(\mathbf{w}_t) \underbrace{-\mathbb{E}_t[\nabla R(\mathbf{w}_t)^\top (D_{t+1}h_{t+1})]}_{\Upsilon} \alpha + \alpha^2 d(1-\eta_2)^{-1}(1-\tau)^{-1}L$$

where $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot|_t]$ implies taking expectation over i_t, ξ given $\mathbf{w}_t$. In the following analysis, we decompose Υ into three parts and bound them one by one:

$$\begin{aligned}
\Upsilon &= -\langle \nabla R(\mathbf{w}_t), D_{t+1}h_{t+1}\rangle = -\langle \nabla R(\mathbf{w}_t), D_t h_{t+1}\rangle - \langle \nabla R(\mathbf{w}_t), (D_{t+1}-D_t)h_{t+1}\rangle \\
&= -(1-\eta_1)\langle \nabla R(\mathbf{w}_t), D_t \nabla g_{i_t}(\mathbf{w}_t;\xi)^\top \nabla f(\mathbf{u}_{i_t})\rangle - \eta_1 \langle \nabla R(\mathbf{w}_t), D_t h_t\rangle \\
&- \langle \nabla R(\mathbf{w}_t), (D_{t+1}-D_t)h_{t+1}\rangle \\
&= I_1^t + I_2^t + I_3^t
\end{aligned}$$

Let us first bound I_1^t,

$$\begin{aligned}
\mathbb{E}_t[I_1^t] &\overset{(a)}{=} -(1-\eta_1)\langle \nabla R(\mathbf{w}_t), \mathbb{E}_t[D_t \nabla g_{i_t}(\mathbf{w}_t;\xi)^\top \nabla f(\mathbf{u}_{i_t})]\rangle \\
&= -(1-\eta_1)\langle \nabla R(\mathbf{w}_t), \mathbb{E}_t[D_t \nabla g_{i_t}(\mathbf{w}_t;\xi)^\top \nabla f(g_{i_t}(\mathbf{w}_t))]\rangle \\
&+ (1-\eta_1)\langle \nabla R(\mathbf{w}_t), \mathbb{E}_t[D_t \nabla g_{i_t}(\mathbf{w}_t;\xi)^\top (\nabla f(\mathbf{u}_{i_t}) - \nabla f(g_{i_t}(\mathbf{w}_t)))]\rangle \\
&\leq -(1-\eta_1)\|\nabla R(\mathbf{w}_t)\|_{D_t}^2 \\
&+ (1-\eta_1)\|D_t^{-1/2}\nabla R(\mathbf{w}_t)\| \mathbb{E}_t[\|D_t^{-1/2}\nabla g_{i_t}(\mathbf{w}_t;\xi)^\top (\nabla f(\mathbf{u}_{i_t}) - \nabla f(g_{i_t}(\mathbf{w}_t)))\|] \\
&\overset{(b)}{\leq} -(1-\eta_1)\|\nabla R(\mathbf{w}_t)\|_{D_t}^2 + \frac{(1-\eta_1)\|\nabla R(\mathbf{w}_t)\|_{D_t}^2}{2} \\
&+ \frac{(1-\eta_1)\mathbb{E}_t[\|D_t^{-1/2}\nabla g_{i_t}(\mathbf{w}_t;\xi)^\top (\nabla f(\mathbf{u}_{i_t}) - \nabla f(g_{i_t}(\mathbf{w}_t)))\|^2]}{2} \\
&\leq -\frac{(1-\eta_1)}{2}\|\nabla R(\mathbf{w}_t)\|_{D_t}^2 + \frac{(1-\eta_1)}{2}\mathbb{E}_t[\|\nabla g_{i_t}(\mathbf{w}_t;\xi)^\top (\nabla f(\mathbf{u}_{i_t}) - \nabla f(g_{i_t}(\mathbf{w}_t))\|_{D_t}^2] \\
&\overset{(c)}{\leq} -\frac{(1-\eta_1)}{2}(\epsilon + C_g^2 C_f^2)^{-1/2}\|\nabla R(\mathbf{w}_t)\|^2 + \frac{1}{2}\epsilon^{-1/2} C_g^2 L_f^2 \mathbb{E}[\|g_{i_t}(\mathbf{w}_t) - \mathbf{u}_{i_t}\|^2]
\end{aligned} \tag{27}$$

where equality (a) is due to $\nabla R(\mathbf{w}_t) = \mathbb{E}_{i_t,\xi}[\nabla g_{i_t}(\mathbf{w}_t;\xi)^\top \nabla f(g_{i_t}(\mathbf{w}_t))]$, where i_t and ξ are independent. The inequality (b) is according to $ab \leq a^2/2 + b^2/2$. The last inequality (c) is due to $\epsilon^{-1/2}\mathbb{I} \geq \|D_t\mathbb{I}\| = \|\frac{1}{\sqrt{\epsilon\mathbb{I} + \hat{v}_{t+1}}}\| \geq \|(\epsilon\mathbb{I} + C_g^2 C_f^2)^{-1/2}\| = (\epsilon + C_g^2 C_f^2)^{-1/2}\mathbb{I}$, $(1-\eta_1) \leq 1$ and

$$\begin{aligned}
&\mathbb{E}_t[\|\nabla g_{i_t}(\mathbf{w}_t;\xi)^\top (\nabla f(\mathbf{u}_{i_t}) - \nabla f(g_{i_t}(\mathbf{w}_t)))\|_{D_t}^2] \\
&\leq \epsilon^{-1/2} C_g^2 \mathbb{E}_t[\|\nabla f(\mathbf{u}_{i_t}) - \nabla f(g_{i_t}(\mathbf{w}_t))\|_{\mathbb{I}}^2] \\
&\leq \epsilon^{-1/2} C_g^2 L_f^2 \mathbb{E}_t[\|g_{i_t}(\mathbf{w}_t) - \mathbf{u}_{i_t}\|^2]
\end{aligned} \tag{28}$$

For I_2^t and I_3^t, we have

$$\begin{aligned}
\mathbb{E}_t[I_2^t] &= -\eta_1\langle \nabla R(\mathbf{w}_t) - \nabla R(\mathbf{w}_{t-1}), D_t h_t\rangle - \eta_1 \langle \nabla R(\mathbf{w}_{t-1}), D_t h_t\rangle \\
&\leq \eta_1 L \alpha^{-1}\|\mathbf{w}_t - \mathbf{w}_{t-1}\|^2 - \eta_1 \langle \nabla R(\mathbf{w}_{t-1}), D_t h_t\rangle \\
&= \eta_1 L \alpha^{-1}\|\mathbf{w}_t - \mathbf{w}_{t-1}\|^2 + \eta_1 (I_1^{t-1} + I_2^{t-1} + I_3^{t-1}) \\
&\leq \eta_1 L \alpha d (1-\eta_2)^{-1}(1-\tau)^{-1} + \eta_1 (I_1^{t-1} + I_2^{t-1} + I_3^{t-1})
\end{aligned} \tag{29}$$

where the last equation applies Lemma 7.

$$\begin{aligned}\mathbb{E}_t[I_3^t] &= -\langle \nabla R(\mathbf{w}_t), (D_{t+1} - D_t)h_{t+1}\rangle = -\sum_{i'=1}^{d} \nabla_{i'} R(\mathbf{w}_t)((\epsilon + \hat{v}_t^{i'})^{-1/2} - (\epsilon + \hat{v}_{t+1}^{i'})^{-1/2})h_{t+1}^{i'} \\ &\leq \|\nabla R(\mathbf{w}_t)\|\|h_{t+1}\| \sum_{i'=1}^{d}((\epsilon + \hat{v}_t^{i'})^{-1/2} - (\epsilon + \hat{v}_{t+1}^{i'})^{-1/2}) \\ &\leq C_g^2 C_f^2 \sum_{i'=1}^{d}((\epsilon + \hat{v}_t^{i'})^{-1/2} - (\epsilon + \hat{v}_{t+1}^{i'})^{-1/2})\end{aligned} \tag{30}$$

By combining Equation (28), (29) and (30) together,

$$\begin{aligned}\mathbb{E}_t[I_1^t + I_2^t + I_3^t] \leq & -\frac{(1-\eta_1)}{2}(\epsilon + C_g^2 C_f^2)^{-1/2}\|\nabla R(\mathbf{w}_t)\|^2 + \frac{1}{2}\epsilon^{-1/2} C_g^2 L_f^2 \mathbb{E}_t[\|g_{i_t}(\mathbf{w}_t) - \mathbf{u}_{i_t}\|^2] \\ & + \eta_1 L\alpha d(1-\eta_2)^{-1}(1-\tau)^{-1} + \eta_1(I_1^{t-1} + I_2^{t-1} + I_3^{t-1}) \\ & + C_g^2 C_f^2 \sum_{i'=1}^{d}((\epsilon + \hat{v}_t^{i'})^{-1/2} - (\epsilon + \hat{v}_{t+1}^{i'})^{-1/2})\end{aligned} \tag{31}$$

Define the Lyapunov function

$$\mathbb{V}_t = R(\mathbf{w}_t) - c_t\langle \nabla R(\mathbf{w}_{t-1}), D_t h_t\rangle \tag{32}$$

where c_t and c will be defined later.

$$\begin{aligned}&\mathbb{E}_t[\mathbb{V}_{t+1} - \mathbb{V}_t] \\ &= R(\mathbf{w}_{t+1}) - R(\mathbf{w}_t) - c_{t+1}\langle \nabla R(\mathbf{w}_t), D_{t+1}h_{t+1}\rangle + c_t\langle \nabla R(\mathbf{w}_{t-1}), D_t h_t\rangle \\ &\leq -(c_{t+1} + \alpha)\langle \nabla R(\mathbf{w}_t), D_{t+1}h_{t+1}\rangle + \frac{L}{2}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2 + c_t\langle \nabla R(\mathbf{w}_{t-1}), D_t h_t\rangle \\ &= (c_{t+1} + \alpha)(I_1^t + I_2^t + I_3^t) + \frac{L}{2}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2 - c_t(I_1^{t-1} + I_2^{t-1} + I_3^{t-1}) \\ &\overset{Eqn\ (31)\ \text{and}\ Lemma\ 7}{\leq} -(\alpha + c_{t+1})\frac{(1-\eta_1)}{2}(\epsilon + C_g^2 C_f^2)^{-1/2}\|\nabla R(\mathbf{w}_t)\|^2 \\ &+ (\alpha + c_{t+1})\eta_1 L\alpha d(1-\eta_2)^{-1}(1-\tau)^{-1} + \eta_1(\alpha + c_{t+1})(I_1^{t-1} + I_2^{t-1} + I_3^{t-1}) \\ &+ (\alpha + c_{t+1})C_g^2 C_f^2 \sum_{i'=1}^{d}((\epsilon + \hat{v}_{i'}^t)^{-1/2} - (\epsilon + \hat{v}_{i'}^{t+1})^{-1/2}) \\ &+ \frac{L}{2}\alpha^2 d(1-\eta_2)^{-1}(1-\tau)^{-1} - c_t(I_1^{t-1} + I_2^{t-1} + I_3^{t-1}) + \frac{\epsilon^{-1/2} C_g^2 L_f^2(\alpha + c_{t+1})}{2}\|g_{i_t}(\mathbf{w}_t) - \mathbf{u}_{i_t}\|^2\end{aligned} \tag{33}$$

By setting $\alpha_{t+1} \leq \alpha_t = \alpha$, $c_t = \sum_{p=t}^{\infty}(\prod_{j=t}^{p}\eta_1)\alpha_j$, and $c = (1 + (1-\eta_1)^{-1})\epsilon^{-\frac{1}{2}} C_g^2 L_f^2$, we have

$$c_t \leq (1-\eta_1)^{-1}\alpha_t,\ \frac{2(\alpha + c_{t+1})}{\alpha}\beta\epsilon^{-1/2} C_g^2 L_f^2 \leq c\beta,\ \eta_1(\alpha + c_{t+1}) = c_t \tag{34}$$

As a result, $\eta_1(\alpha + c_{t+1})(I_1^{t-1} + I_2^{t-1} + I_3^{t-1}) - c_t(I_1^{t-1} + I_2^{t-1} + I_3^{t-1}) = 0$

$$\begin{aligned}
\mathbb{E}_t[\mathbb{V}_{t+1} - \mathbb{V}_t] &\leq -(\alpha + c_{t+1})\frac{(1-\eta_1)}{2}(\epsilon + C_g^2 C_f^2)^{-1/2}\|\nabla R(\mathbf{w}_t)\|^2 \\
&+ (\alpha + c_{t+1})\eta_1 L\alpha d(1-\eta_2)^{-1}(1-\tau)^{-1} + \frac{L}{2}\alpha^2 d(1-\eta_2)^{-1}(1-\tau)^{-1} \\
&+ (\alpha + c_{t+1})C_g^2 C_f^2 \sum_{i'=1}^{d}((\epsilon + \hat{v}_{i'}^t)^{-1/2} - (\epsilon + \hat{v}_{i'}^{t+1})^{-1/2}) \\
&+ \frac{(\alpha + c_{t+1})}{2}\epsilon^{-1/2} C_g^2 L_f^2 \|g_{i_t}(\mathbf{w}_t) - \mathbf{u}_{i_t}\|^2 \\
&\leq -\alpha\frac{(1-\eta_1)}{2}(\epsilon + C_g^2 C_f^2)^{-1/2}\|\nabla R(\mathbf{w}_t)\|^2 \\
&+ 2\eta_1 L\alpha^2 T d(1-\eta_1)^{-1}(1-\eta_2)^{-1}(1-\tau)^{-1} + \frac{L}{2}T\alpha^2 d(1-\eta_2)^{-1}(1-\tau)^{-1} \\
&+ 2(1-\eta_1)^{-1}\alpha C_g^2 C_f^2 \sum_{i'=1}^{d}((\epsilon + \hat{v}_{i'}^t)^{-1/2} - (\epsilon + \hat{v}_{i'}^{t+1})^{-1/2}) + \frac{c\alpha}{4}\sum_{t=1}^{T}\mathbb{E}_t[\|g_{i_t}(\mathbf{w}_t) - \mathbf{u}_{i_t}\|^2]
\end{aligned} \tag{35}$$

where the last inequality is due to equation (34) such that we have $2(\alpha + c_{t+1})\epsilon^{-1/2} C_g^2 L_f^2 \leq c\alpha$, and $\alpha + c_{t+1} \leq 2(1-\eta_1)^{-1}\alpha$.
Then by rearranging terms, and taking summation from $1, \cdots, T$ of equation (35), we have

$$\begin{aligned}
&\sum_{t=1}^{T}\alpha\frac{(1-\eta_1)}{2}(\epsilon + C_g^2 C_f^2)^{-1/2}\|\nabla R(\mathbf{w}_t)\|^2 \leq \sum_{t=1}^{T}\mathbb{E}_t[\mathbb{V}_t - \mathbb{V}_{t+1}] \\
&+ 2\eta_1 L\alpha^2 T d(1-\eta_1)^{-1}(1-\eta_2)^{-1}(1-\tau)^{-1} + LT\alpha^2 d(1-\eta_2)^{-1}(1-\tau)^{-1} \\
&+ 2(1-\eta_1)^{-1}\alpha C_g^2 C_f^2 \sum_{t=1}^{T}\sum_{i'=1}^{d}((\epsilon + \hat{v}_{i'}^t)^{-1/2} - (\epsilon + \hat{v}_{i'}^{t+1})^{-1/2}) + c\alpha\sum_{t=1}^{T}\mathbb{E}_t[\|g_{i_t}(\mathbf{w}_t) - \mathbf{u}_{i_t}\|^2] \\
&\leq \mathbb{E}[\mathbb{V}_1] - \mathbb{E}[\mathbb{V}_{T+1}] \\
&+ 2\eta_1 L\alpha^2 T d(1-\eta_1)^{-1}(1-\eta_2)^{-1}(1-\tau)^{-1} + LT\alpha^2 d(1-\eta_2)^{-1}(1-\tau)^{-1} \\
&+ 2(1-\eta_1)^{-1}\alpha C_g^2 C_f^2 \sum_{i'=1}^{d}((\epsilon + \hat{v}_0^{i'})^{-1/2}) + c\alpha\sum_{t=1}^{T}\mathbb{E}_t[\|g_{i_t}(\mathbf{w}_t) - \mathbf{u}_{i_t}\|^2]
\end{aligned} \tag{36}$$

By combing with Lemma 4, We finish the proof.