RDA: Reciprocal Distribution Alignment for Robust Semi-supervised Learning

Yue Duan[1], Lei Qi[2], Lei Wang[3], Luping Zhou[4], and Yinghuan Shi[1]

[1] Nanjing University, China [2] Southeast University, China
[3] University of Wollongong, Australia [4] University of Sydney, Australia

A Proof of Theorem 1

In this section, we consider the case where $p_1 \geq \frac{1}{2}$ for the proof of Theorem 1.

Proof. Rewriting Eq. (15), we obtain

$$\sum_{i=1}^{n} p_i \log p_i \geq \sum_{i=1}^{n} \overline{p}_i \log \overline{p}_i. \tag{22}$$

Denoting $\overline{p}_1 = \frac{1-p_n}{n-1}, \ldots, \overline{p}_n = \frac{1-p_1}{n-1}$, we have

$$\frac{1}{n-1} \geq \overline{p}_1 \geq \cdots \geq \overline{p}_n. \tag{23}$$

Let $\boldsymbol{a} = (\overline{p}_1, \ldots, \overline{p}_{n-1}, \overline{p}_n)$ and $\boldsymbol{b} = (\frac{1}{n-1}, \ldots, \frac{1}{n-1}, 0)$, by Eq. (23) and $\sum_{i=1}^{n} \overline{p}_i = \sum_{i=1}^{n-1} \frac{1}{n-1} = 1$, we notice $\boldsymbol{a}$ is majorized by $\boldsymbol{b}$ ($\boldsymbol{a} \prec \boldsymbol{b}$) [8,1]. Since the function $g(\boldsymbol{x}) = \sum_{i=1}^{d} x_i \log(x_i)$ is Schur-convex [10,11], we have $g(\boldsymbol{a}) \leq g(\boldsymbol{b})$ [10,11], *i.e.*,

$$\sum_{i=1}^{n} \overline{p}_i \log \overline{p}_i \leq (n-1)\frac{1}{n-1} \log \frac{1}{n-1} = -\log(n-1). \tag{24}$$

Next, rewriting the left term in Eq. (22), we have

$$\sum_{i=1}^{n} p_i \log p_i = p_1 \log p_1 + \sum_{i=2}^{n} p_i \log p_i. \tag{25}$$

Since $p_2 + \cdots + p_n = 1 - p_1$ and $g(x) = x \log x$ is a convex function, by Jensen's Inequality, we obtain the minimum of $\sum_{i=2}^{n} p_i \log p_i$ when $p_2 = \cdots = p_n = \frac{1-p_1}{n-1}$. Then, by Eq. (25), we have

$$\begin{aligned}
\sum_{i=1}^{n} p_i \log p_i &\geq p_1 \log p_1 + (\frac{1-p_1}{n-1} \log \frac{1-p_1}{n-1})(n-1) \\
&= p_1 \log p_1 + (1-p_1)\log(1-p_1) - (1-p_1)\log(n-1) \\
&\geq -1 - \frac{1}{2}\log(n-1). \\
&\quad \text{(using } p_1 \log p_1 + (1-p_1)\log(1-p_1) \geq -\log 2 \text{ and } 1-p_1 \leq \tfrac{1}{2}\text{)}
\end{aligned}$$

Notice that by Eq. (24) we have $\sum_{i=1}^{n} \overline{p}_i \log \overline{p}_i \leq -\log(n-1)$. Solving inequality

$$-1 - \frac{1}{2}\log(n-1) \geq -\log(n-1), \tag{26}$$

we obtain $n \geq 5$. Theorem 1 now follows simply by combining proofs for the cases where $p_1 < \frac{1}{2}$ and $p_1 \geq \frac{1}{2}$. To sum up, for multi-classification tasks, we prove that when $n \geq 5$, $\mathcal{H}(\overline{p}) \geq \mathcal{H}(p)$ holds, *i.e.*, Reverse Operation could maximize the entropy of p. The proof for complementary label version can be simply obtained by replacing p and $\overline{p}$ in the above formulas with q and $\overline{q}$, respectively. □

B Algorithm

Pseudo-code of RDA is shown in Algorithm 1.

Algorithm 1: RDA: Reciprocal Distribution Alignment

Input: batch of labeled data $\mathcal{X} = \{(x_b, y_b)\}_{b=1}^{B}$, batch of unlabeled data $\mathcal{U} = \{u_b\}_{b=1}^{\mu B}$, Default Classifier $\mathcal{D}$, Auxiliary Classifier $\mathcal{A}$, maximum number of iterations M, augmentation α

for *iteration* $t = 1$ *to* M **do**
  // Select complementary label from $\mathcal{Y}$ randomly
  $\overline{y}_b = \text{randselect}(\mathcal{Y} \setminus \{y_b\}), b \in (1, \ldots, B)$
  // Compute default supervised loss
  $\mathcal{L}_{sd} = \frac{1}{B}\sum_{n=1}^{B} H(y_n, P_{\mathcal{D}}(y_c|x_{w,n}))$
  // Compute auxiliary supervised loss
  $\mathcal{L}_{sa} = \frac{1}{B}\sum_{n=1}^{B} H(\overline{y}_n, P_{\mathcal{A}}(y_c|x_{w,n}))$
  for *iteration* $b = 1$ *to* μB **do**
    $u_{w,b} = \alpha_{\text{weak}}(u_b)$ // Apply weak augmentation to u_b
    $u_{s,b} = \alpha_{\text{strong}}(u_b)$ // Apply strong augmentation to u_b
    $p_b = P_{\mathcal{D}}(y_c|u_{w,b})$ // Compute predictions of $\mathcal{D}$ for $u_{w,b}$
    $p_{s,b} = P_{\mathcal{D}}(y_c|u_{s,b})$ // Compute predictions of $\mathcal{D}$ for $u_{s,b}$
    $q_b = P_{\mathcal{A}}(y_c|u_{w,b})$ // Compute predictions of $\mathcal{A}$ for $u_{w,b}$
    $q_{s,b} = P_{\mathcal{A}}(y_c|u_{s,b})$ // Compute predictions of $\mathcal{A}$ for $u_{s,b}$
    $\overline{p}_b = \text{Norm}(\mathbb{1} - p_b)$
    $\overline{q}_b = \text{Norm}(\mathbb{1} - q_b)$
    // Apply distribution alignment reciprocally
    $\tilde{p}_b = \text{Norm}(p_b \times \frac{\Psi(\overline{q})}{\Psi(p)})$
    $\tilde{q}_b = \text{Norm}(q_b \times \frac{\Psi(\overline{p})}{\Psi(q)})$ // Soft complementary labels for $u_{w,b}$
    $\hat{\tilde{p}}_b = \arg\max(\tilde{p}_b)$ // Hard pseudo-labels for $u_{w,b}$
  end
  $\mathcal{L}_{cd} = \frac{1}{\mu B}\sum_{n=1}^{\mu B} H(\hat{\tilde{p}}_n, p_{s,n})$ // Compute default consistency loss
  $\mathcal{L}_{ca} = \frac{1}{\mu B}\sum_{n=1}^{\mu B} H(\tilde{q}_n, q_{s,n})$ // Compute auxiliary consistency loss
  return $\mathcal{L} = \mathcal{L}_{sd} + \lambda_a \mathcal{L}_{sa} + \lambda_{cd}\mathcal{L}_{cd} + \lambda_{ca}\mathcal{L}_{ca}$ // Optimize total loss $\mathcal{L}$
end

Table 7. More baseline results in accuracy (%) under DARP's protocol. Results of baseline methods are copied from DAPR [4].

Method	CIFAR-10 (γ_l =100)				STL-10 ($\gamma_l \neq \gamma_u$)	
	$\gamma_u = 1$	$\gamma_u = 50$	$\gamma_u = 150$	$\gamma_u = 100$ (reversed)	$\gamma_l = 10$	$\gamma_l = 20$
MixMatch	41.50±0.76	64.10±0.58	65.50±0.64	47.90±0.09	56.30±0.46	45.20±0.19
M w. DARP	86.70±0.80	68.30±0.47	66.70±0.25	72.90±0.24	67.90±0.24	58.30±0.73
ReMixMatch	48.30±0.14	75.10±0.43	72.50±0.10	49.00±0.55	67.80±0.45	60.10±1.18
R w. DARP	89.70±0.15	77.40±0.22	73.20±0.11	**80.10±0.11**	79.40±0.07	70.90±0.44
RDA	**93.35±0.24**	**79.77±0.06**	**74.48±0.24**	79.25±0.52	**87.21±0.44**	**83.21±0.52**

C Datasets with Mismatched distributions

C.1 Protocol of DARP

DARP [4] introduces this protocol to build a class-imbalanced dataset. DARP introduces two parameters namely imbalanced ratio γ_l and γ_u to control the class-imbalance of dataset. For the labeled data, the data number of each class N_i is scaled by: $N_i = N_1 \times \gamma_l^{-\frac{i-1}{n-1}}$, where $i \in (1, \dots, n)$ and n is the number of classes. Likewise, for the unlabeled data, the data number of each class M_i is scaled by: $M_i = M_1 \times \gamma_u^{-\frac{i-1}{n-1}}$. Specially, *"reversed"* in Tab. 5 indicates that the unlabeled data with reversely ordered class distribution is used, *i.e.*, $M_i = M_1 \times \gamma_u^{-\frac{n-i}{n-1}}$. $N_1 = 1500$ and $M_1 = 3000$ are applied into CIFAR-10 under DARP's protocol. DARP constructs STL-10 with $N_1 = 450$ and fully use the given unlabeled data in this dataset (*i.e.*, $\sum_{i=1}^{n} M_i = 100,000$). γ_u is not set for STL-10 due to the unknown ground-truth of the unlabeled data. DARP claims the labeled and unlabeled data in STL-10 have different distributions, *i.e.*, $\gamma_l \neq \gamma_u$.

Additionally, we show the results of more baseline methods under DARP's protocol [4] in Tab. 7 for comparison with our method.

C.2 Imbalanced C_x

We now show the details on how to construct dataset with imbalanced labeled data (*i.e.*, C_x is imbalanced) while keeping the number of labeled data unchanged. Following CIFAR-LT [2], we mimic the imbalanced C_x by an exponential function: $N_i = N_0 \times \gamma_x^{-\frac{i-1}{n-1}}, i \in (1, \dots n)$ to generate the number of labeled data for class with index i, where n is the number of classes. We use different N_0 to investigate different scale of imbalance. With N_0 we set, γ_x is calculated by the constraint $\sum_{i=1}^{n} N_i = D_x$, where D_x is the number of labels we set. We search for a γ_x from small to large in natural numbers, so that the progress of search can be summarized as the following optimization:

$$\begin{aligned} \hat{\gamma_x} = \arg\min_{\gamma_x} \quad & D_x - \sum_{i=1}^{n} N_i \\ s.t. \quad & D_x - \textstyle\sum_{i=1}^{n} N_i > 0 \end{aligned} \tag{27}$$

With obtained γ_x, we add missing labels for classes other than the first class (*i.e.*, keep the N_0 unchanged) in turn until the condition $\sum_{i=1}^{n} N_i = D_x$ is met. Here we found that the labels that need to be added are less than n, which means we can complete this progress by adding at most one round in turn.

D Additional Experiments with Mismatched Distributions

D.1 Mismatched Distributions with Non-overlapping Classes in the Unlabeled Data

In addition to the mismatched distributions discussed in Sec. 3.1, SSL with non-overlapping classes in the unlabeled data is a more generalized mismatched scenario. As mentioned, this distribution mismatch is known as SSL using *out-of-distribution* **(OOD)** samples in the unlabeled data [9] (also known as *open-set SSL*). To explore the robustness of RDA, we experiment under the same setting as Sec. 4.4 in [9] and observe slight accuracy drops of RDA, except for at 100% class mismatch extent (sometimes more than 10% drop). This is understandable because SSL with OOD samples is very different from our task addressing the mismatched distributions with the same classes and we learn total unlabeled data without OOD sample filters. Considering the fine-grained datasets *Semi-Aves* [14] and *Semi-Fungi* [13] are also used to mimic the OOD setting [13], we evaluate our RDA on them. As shown in Tab. 8, when suffers from both mismatched distributions (in our paper) and OOD samples, RDA can still outperform our main baseline FixMatch by improving the pseudo-labels with in-distribution classes, although some aligned pseudo-labels may be assigned to OOD samples. In the future, we will extend RDA to handle open-set SSL, *e.g.*, detecting OOD samples from the perspective of distribution. Furthermore, we provide discussions on the mismatched distributions with completely disjoint classes in C_x and C_u. This scenario is an extreme case to SSL with OOD samples and *few-label transfer* proposed in [7] is closely related to it. Differently, our paper argues that even in the normal SSL setting where C_x and C_u share the same classes, the mismatched distributions could cause significant degradation of many popular SSL methods. Considering RDA is originally designed to strategically align distributions of overlapping classes, it could not work with completely disjoint C_x and C_u.

D.2 Learning with Symmetric Noisy Labels

This is a novel setting different from the previous mismatched setting. We note that there are some subtle connections between dataset with noise and mismatched distributions dataset. We treat the total data in the dataset with noise as labeled data and also treat them as unlabeled data, *i.e.*, this scenario can be seen as a process of SSL. Asymmetric noise is designed by mapping ground-truth labels to similar classes. *e.g.*, in CIFAR-10, we generate noisy labels by deer→ horse, dog↔ cat, *etc.* Thus, we can regard CIFAR-10 with asymmetric noise as a mismatched dataset, *i.e.*, the existence of asymmetric noise increases the ratio of some classes and decreases the ratio of some classes accordingly.

Table 8. Accuracy (%) in open-set SSL. Both Semi-Aves and Sem-Fungi have not only OOD unlabeled data but also in-distribution unlabeled data within class distribution that mismatches with the labeled data [13]. Unlike native RDA, we revert to the confidence-based thresholding in FixMatch [12] as a simple filter for OOD samples. While this goes against our original intention of using only distribution alignment to improve pseudo-labeling, it is a compromise for this open-set scenario. We follow the backbone and hyper-parameters for FixMatch in [13] and train models from scratch.

Method	Semi-Aves	Semi-Fungi
	Top-1 / Top-5	Top-1 / Top-5
FixMatch	19.2 / 42.6	25.2 / 50.2
RDA	**21.9** / **43.7**	**28.7** / **51.2**

Table 9. Results of accuracy (%) on CIFAR-10 using full labels with 40% asymmetric noise. Results of baseline noisy label learning methods are reported in DivideMix [5].

Method	CIFAR-10
	40% asym noise
P-correction [16]	88.5
Joint-Optim [15]	88.9
Meta-Learning [6]	89.2
DivideMix [5]	**93.4**
RDA	90.5

We evaluate RDA on CIFAR-10 with 40% asymmetric noise. Following DivideMix [5], the backbone used in experiments is 18-layer PreAct ResNet [3] and we train the models with the same setting in Sec. 4.4. Although we do not make a special design for noisy label, RDA still achieves quite competitive performance compared with the noisy label learning methods shown in Tab. 9.

References

1. Arnold, B.C.: Majorization and the Lorenz order: A brief introduction, vol. 43. Springer Science & Business Media (2012)
2. Cao, K., Wei, C., Gaidon, A., Arechiga, N., Ma, T.: Learning imbalanced datasets with label-distribution-aware margin loss. arXiv preprint arXiv:1906.07413 (2019)
3. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: European Conference on Computer Vision (2016)
4. Kim, J., Hur, Y., Park, S., Yang, E., Hwang, S.J., Shin, J.: Distribution aligning refinery of pseudo-label for imbalanced semi-supervised learning. In: Advances in Neural Information Processing Systems (2020)
5. Li, J., Socher, R., Hoi, S.C.: Dividemix: Learning with noisy labels as semi-supervised learning. In: International Conference on Learning Representations (2020)
6. Li, J., Wong, Y., Zhao, Q., Kankanhalli, M.S.: Learning to learn from noisy labeled data. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019)

7. Li, S., Chen, D., Chen, Y., Yuan, L., Zhang, L., Chu, Q., Liu, B., Yu, N.: Improve unsupervised pretraining for few-label transfer. In: IEEE/CVF International Conference on Computer Vision. pp. 10201–10210 (2021)
8. Marshall, A.W., Olkin, I., Arnold, B.C.: Inequalities: theory of majorization and its applications, vol. 143. Springer (1979)
9. Oliver, A., Odena, A., Raffel, C.A., Cubuk, E.D., Goodfellow, I.: Realistic evaluation of deep semi-supervised learning algorithms. In: Advances in Neural Information Processing Systems (2018)
10. Peajcariaac, J.E., Tong, Y.L.: Convex functions, partial orderings, and statistical applications. Academic Press (1992)
11. Roberts, A.W.: Convex functions. In: Handbook of convex geometry, pp. 1081–1104. Elsevier (1993)
12. Sohn, K., Berthelot, D., Li, C.L., Zhang, Z., Carlini, N., Cubuk, E.D., Kurakin, A., Zhang, H., Raffel, C.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In: Advances in Neural Information Processing Systems (2020)
13. Su, J.C., Cheng, Z., Maji, S.: A realistic evaluation of semi-supervised learning for fine-grained classification. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021)
14. Su, J.C., Maji, S.: The semi-supervised inaturalist-aves challenge at fgvc7 workshop. arXiv preprint arXiv:2103.06937 (2021)
15. Tanaka, D., Ikami, D., Yamasaki, T., Aizawa, K.: Joint optimization framework for learning with noisy labels. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018)
16. Yi, K., Wu, J.: Probabilistic end-to-end noise correction for learning with noisy labels. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019)