A. Deferred proofs

Proof of Lemma 1. For the sake of the proof, assume that we have two labeled set of samples of size m and m_L from $p_{\mathcal{XY}}$, call them respectively S and S_L. The set S represents our unlabeled sample, and the set S_L represents the labeled sample. For any $\delta \in (0,1)$, we would like to find a $\gamma > 0$ such that with probability $1-\delta$, for all $i \in 1, \dots, n$,

$$\left| \frac{1}{m} \sum_{(x,\boldsymbol{y}) \in S} \ell(\boldsymbol{\phi}_i(x), \boldsymbol{y}) - \frac{1}{m_L} \sum_{(x,\boldsymbol{y}) \in S_L} \ell(\boldsymbol{\phi}_i(x), \boldsymbol{y}) \right| \leq \gamma \ . \tag{6}$$

The sample S represents the unlabeled data $x_1, \dots, x_m$ we have access to. In fact, $\frac{1}{m} \sum_{(x,\boldsymbol{y}) \in S} \ell(\boldsymbol{\phi}_i(x), \boldsymbol{y}) = \hat{R}(\boldsymbol{\phi}_i; X, \boldsymbol{Y}^*)$. The inequality (6) implies that for the true labeling of the unlabeled data $x_1, \dots, x_m$, for any $i \in 1, \dots, n$, it holds that:

$$\hat{R}(\boldsymbol{\phi}_i; X, \boldsymbol{Y}^*) \in [\hat{\mu}_i - \gamma, \hat{\mu}_i + \gamma]$$

where $\hat{\mu}_i = \frac{1}{m_L} \sum_{(x,\boldsymbol{y}) \in S_L} \ell(\boldsymbol{\phi}_i(x), \boldsymbol{y})$ is the empirical mean computed from the labeled sample S_L.

By using Hoeffding's inequality, we have that for a fixed i, it holds that

$$\begin{aligned} &\mathop{\mathbb{P}}_{S,S_L} \left(\left| \tfrac{1}{m} \sum_{(x,\boldsymbol{y}) \in S} \ell(\boldsymbol{\phi}_i(x), \boldsymbol{y}) - \tfrac{1}{m_L} \sum_{(x,\boldsymbol{y}) \in S_L} \ell(\boldsymbol{\phi}_i(x), \boldsymbol{y}) \right| > \gamma \right) \\ &\leq 2 \exp\left(\frac{-2\gamma^2}{\sum_{j=1}^{m} (\frac{B}{m})^2 + \sum_{j=1}^{m_L} (\frac{B}{m_L})^2} \right) \\ &= 2 \exp\left(\frac{-2 m_L m \gamma^2}{B^2 (m + m_L)} \right) = \frac{\delta}{n} \ . \end{aligned} \tag{7}$$

By taking a union bound and solving (7) with respect to γ, the statement follows.

Proof of Lemma 4. By invoking Lemma 3, it is easy to see that the function $\hat{R}(\boldsymbol{h_\theta}; X, \boldsymbol{Y'})$ is a convex combination of convex functions with respect to $\boldsymbol{\theta}$, hence it is also convex in $\boldsymbol{\theta}$. Let $\boldsymbol{v} \in \partial \hat{R}(\boldsymbol{h_\theta}; X, \boldsymbol{Y'})$. If a function is convex, then there exists at least one subgradient for each point of its domain, so $\boldsymbol{v}$ is well defined. Then, we have that for any $\boldsymbol{\theta''} \in \Theta$, it holds that

$$\hat{R}(\boldsymbol{h_{\theta''}}, \boldsymbol{Y'}) - \hat{R}(\boldsymbol{h_{\theta'}}, \boldsymbol{Y'}) \geq \boldsymbol{v}^T (\boldsymbol{\theta''} - \boldsymbol{\theta'}) \ .$$

As $f(\boldsymbol{\theta''}) \geq \hat{R}(\boldsymbol{h_{\theta''}}; X, \boldsymbol{Y'})$, we have that

$$f(\boldsymbol{\theta''}) - f(\boldsymbol{\theta'}) \geq \boldsymbol{v}^T (\boldsymbol{\theta''} - \boldsymbol{\theta'}) \ ,$$

which implies that $\boldsymbol{v}$ is a subgradient of f at $\boldsymbol{\theta'}$.

Proof of Theorem 5. We need to show that $f(\boldsymbol{\theta})$ is convex and L-Lipschitz continuous with respect to $\boldsymbol{\theta}$ to apply the standard convergence result for constant step size subgradient optimization (Bertsekas, 2015), which yields

$$f(\tilde{\boldsymbol{\theta}}) - f(\hat{\boldsymbol{\theta}}) \leq \frac{\text{diameter}(\Theta)^2 + L^2 h^2 T}{2hT} \ . \tag{8}$$

To show that $f(\boldsymbol{\theta})$ is convex it is straightforward to see that $\hat{R}(\boldsymbol{h_\theta}; X, \boldsymbol{Y})$ is convex in $\boldsymbol{\theta}$ as it is the convex combination of convex functions in $\boldsymbol{\theta}$. For any $\lambda \in [0,1]$, we have that

$$\begin{aligned} f(\lambda \boldsymbol{\theta}_1 + (1-\lambda)\boldsymbol{\theta}_2) &= \max_{\boldsymbol{Y} \in \mathbb{Y}^\diamond} \hat{R}(h_{\lambda \boldsymbol{\theta}_1 + (1-\lambda)\boldsymbol{\theta}_2}; X, \boldsymbol{Y}) \\ &\leq \max_{\boldsymbol{Y} \in \mathbb{Y}^\diamond} \left[\lambda \hat{R}(\boldsymbol{h_{\theta_1}}; X, \boldsymbol{Y}) + (1-\lambda) \hat{R}(\boldsymbol{h_{\theta_2}}; X, \boldsymbol{Y}) \right] \\ &\leq \lambda \max_{\boldsymbol{Y} \in \mathbb{Y}^\diamond} \hat{R}(\boldsymbol{h_{\theta_1}}; X, \boldsymbol{Y}) + (1-\lambda) \max_{\boldsymbol{Y} \in \mathbb{Y}^\diamond} \hat{R}(\boldsymbol{h_{\theta_2}}; X, \boldsymbol{Y}) \\ &= \lambda f(\boldsymbol{\theta}_1) + (1-\lambda) f(\boldsymbol{\theta}_2) \ . \end{aligned}$$

Also, $f(\boldsymbol{\theta})$ is L-Lipschitz continuous with respect to $\boldsymbol{\theta}$. In fact, it is straightforward to see that $\hat{R}(\boldsymbol{h_\theta}; X, \boldsymbol{Y})$ is also L-Lipschitz continuous with respect to $\boldsymbol{\theta}$. For any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$, we have that

$$\begin{aligned} |f(\boldsymbol{\theta}_1) - f(\boldsymbol{\theta}_2)| &\leq \max_{\boldsymbol{Y} \in \mathbb{Y}^\diamond} |\hat{R}(\boldsymbol{h_{\theta_1}}; X, \boldsymbol{Y}) - \hat{R}(\boldsymbol{h_{\theta_2}}; X, \boldsymbol{Y})| \\ &\leq L ||\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2||_2 \ . \end{aligned}$$

The subgradient of $f(\boldsymbol{\theta})$ in $\boldsymbol{\theta}$ is computed by using Lemma 4. The last part of the Theorem immediately follows by substituting h and T in (8) as in the Theorem statement.

Proof of Lemma 6. For any $i \in 1, \dots, n$, we have that

$$\frac{\partial}{\partial \theta_i} \ell(\boldsymbol{h_\theta}(x), \boldsymbol{e}) = 2 \left(\boldsymbol{\phi}_i(x)^T \cdot \boldsymbol{h_\theta}(x) - \boldsymbol{\phi}_i(x)^T \cdot \boldsymbol{e} \right) \ .$$

Therefore, we can bound the norm of the gradient of ℓ as

$$\begin{aligned} ||\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{h_\theta}(x), \boldsymbol{e})||_2 &= 2 \sqrt{\sum_{i=1}^{n} \left(\boldsymbol{\phi}_i(x)^T \cdot (\boldsymbol{h_\theta}(x) - \boldsymbol{e}) \right)^2} \\ &\leq 2 \sqrt{\sum_{i=1}^{n} (1)^2} \\ &\leq 2\sqrt{n} \ . \end{aligned}$$

The first inequality is an application of Hölder's Inequality, as $\left\| \boldsymbol{\phi}_i(x)^T \right\|_1 = 1$ and $\left\| \boldsymbol{h_\theta}(x) - \boldsymbol{e} \right\|_\infty \leq 1$. This implies that the function $\ell(\boldsymbol{h_\theta}(x), \boldsymbol{e})$ is $2\sqrt{n}$-Lipschitz continuous with respect to $\boldsymbol{\theta}$.

Proof of Lemma 7. First, we will prove that $\ell(\boldsymbol{h_\theta}(\boldsymbol{x}), \boldsymbol{e})$ is bounded. Without loss of generality, suppose that $e_i = 1$. We have that

$$\ell(\boldsymbol{h_\theta}(\boldsymbol{x}), \boldsymbol{e}) = -\ln \left(\frac{\exp(\boldsymbol{w}_i^T \cdot \boldsymbol{x})}{\sum_{c=1}^{k} \exp(\boldsymbol{w}_c^T \cdot \boldsymbol{x})} \right) \ .$$

It is easy to see that $\ell(\boldsymbol{h_\theta}(\boldsymbol{x}), \boldsymbol{e}) \geq 0$. By using the Cauchy-Schwarz inequality, we have that

$$\begin{aligned}\ell(\boldsymbol{h_\theta}(\boldsymbol{x}), \boldsymbol{e}) &= -\ln\left(\frac{\exp(\boldsymbol{w}_i^T \cdot \boldsymbol{x})}{\sum_{c=1}^k \exp(\boldsymbol{w}_c^T \cdot \boldsymbol{x})}\right) \\ &\leq -\ln\left(\frac{\exp(-B_w B_x)}{k\exp(B_w B_x)}\right) \\ &\leq 2B_w B_x + \ln k \ .\end{aligned}$$

Now, we prove that $\ell(\boldsymbol{h_\theta}(\boldsymbol{x}), \boldsymbol{e})$ is Lipschitz continuous with respect to $\boldsymbol{\theta}$. For a fixed $(\boldsymbol{x}, \boldsymbol{e}) \in \mathcal{X} \times \mathcal{Y}$, consider the function $\omega(\boldsymbol{p}) : \mathbb{R}^k \to \mathcal{Y}^\diamond$, defined as

$$\omega(\boldsymbol{p}) \doteq -\sum_{c=1}^k e_c \cdot \ln(p_c) \ ,$$

and let

$$\boldsymbol{h}(\boldsymbol{\theta}) \doteq \left(\frac{\exp(\boldsymbol{w}_1^T \cdot \boldsymbol{x})}{\sum_{c=1}^k \exp(\boldsymbol{w}_c^T \cdot \boldsymbol{x})}, \ldots, \frac{\exp(\boldsymbol{w}_k^T \cdot \boldsymbol{x})}{\sum_{c=1}^k \exp(\boldsymbol{w}_c^T \cdot \boldsymbol{x})}\right)^T ,$$

where $\boldsymbol{\theta} = (\boldsymbol{w_1} \ldots \boldsymbol{w_k})^T$, and observe that $\ell(\boldsymbol{h}_\theta(x), \boldsymbol{e}) = \omega \circ \boldsymbol{h}(\boldsymbol{\theta})$.

It is well known that ℓ is $L_\omega L_{\boldsymbol{h}}$-Lipschitz continuous with respect to $\boldsymbol{\theta}$, where L_ω and $L_{\boldsymbol{h}}$ are the Lipschitz constants respectively of ω and $\boldsymbol{h}$. It is also a known result that $L_\omega \leq 1$ (see for example Proposition 4 of (Gao & Pavel, 2018)).

We now want to compute $L_{\boldsymbol{h}}$. We will use the fact that $\max_{\boldsymbol{\theta}\in\Theta} ||J_{\boldsymbol{h}}(\boldsymbol{\theta})||_F \leq L_{\boldsymbol{h}}$, where $J_{\boldsymbol{h}}$ denotes the Jacobian matrix of $\boldsymbol{h}$ and $||\cdot||_F$ denotes the Frobenius norm.

For ease of notation, let $\boldsymbol{h}(\boldsymbol{\theta}) = \boldsymbol{p} = (p_1, \ldots, p_k)^T$. We have that for any $i \in 1, \ldots, k$, it holds that

$$\begin{aligned}\frac{\partial[\boldsymbol{h}(\boldsymbol{\theta})]_i}{\partial \boldsymbol{w}_j} &= p_i p_j \boldsymbol{x} \quad \text{for } j \neq i \ , \text{ and} \\ \frac{\partial[\boldsymbol{h}(\boldsymbol{\theta})]_i}{\partial \boldsymbol{w}_i} &= (p_i - p_i^2)\boldsymbol{x} \ .\end{aligned}$$

Therefore, we can bound the square of the Frobenius norm of the Jacobian matrix of $\boldsymbol{h}$ with

$$\begin{aligned}\left\|J_{\boldsymbol{h}}(\boldsymbol{\theta})\right\|_F^2 &= \sum_{i,j}\left\|\frac{\partial[\boldsymbol{h}(\boldsymbol{\theta})]_i}{\partial \boldsymbol{w}_j}\right\|_2^2 \\ &\leq ||\boldsymbol{x}||_2^2\left(\sum_i [p_i(1-p_i)]^2 + \sum_{i\neq j}[p_i p_j]^2\right) \\ &\leq ||\boldsymbol{x}||_2^2(k + k^2/2) \leq ||k\boldsymbol{x}||_2^2 \ .\end{aligned}$$

We can conclude that $\boldsymbol{h}$ is kB_x-Lipschitz continuous, and the statement follows.

Proof of Theorem 8. From Chapter 14 of Mitzenmacher & Upfal (2017), we know that

$$R(\boldsymbol{h}_{\hat{\boldsymbol{\theta}}}) \leq \hat{R}(\boldsymbol{h}_{\hat{\boldsymbol{\theta}}}; X, \boldsymbol{Y}^*) + 2\hat{\mathfrak{R}}_m(\mathcal{L}; X, \boldsymbol{Y}^*) + O\left(B\sqrt{\frac{\ln\frac{1}{\delta}}{m}}\right) \ .$$

By definition of $f(\cdot)$, it holds that $\hat{R}(\boldsymbol{h}_{\hat{\boldsymbol{\theta}}}; X, \boldsymbol{Y}^*) \leq f(\hat{\boldsymbol{\theta}})$. As $\hat{\boldsymbol{\theta}}$ is the optimal solution of (2), we have that $f(\hat{\boldsymbol{\theta}}) \leq f(\boldsymbol{\theta^*})$. Let $\boldsymbol{Y}' \doteq \arg\max_{\boldsymbol{Y}\in\mathbb{Y}^\diamond} \hat{R}(\boldsymbol{h}_{\boldsymbol{\theta}^*}; X, \boldsymbol{Y})$. It holds that

$$\begin{aligned}&f(\boldsymbol{\theta^*}) = \hat{R}(\boldsymbol{h}_{\boldsymbol{\theta}^*}; X, \boldsymbol{Y}') \\ &= \hat{R}(\boldsymbol{h}_{\boldsymbol{\theta}^*}; X, \boldsymbol{Y}') + \hat{R}(\boldsymbol{h}_{\boldsymbol{\theta}^*}; X, \boldsymbol{Y}^*) - \hat{R}(\boldsymbol{h}_{\boldsymbol{\theta}^*}; X, \boldsymbol{Y}^*) \\ &= \hat{R}(\boldsymbol{h}_{\boldsymbol{\theta}^*}; X, \boldsymbol{Y}^*) + |\hat{R}(\boldsymbol{h}_{\boldsymbol{\theta}^*}; X, \boldsymbol{Y}') - \hat{R}(\boldsymbol{h}_{\boldsymbol{\theta}^*}; X, \boldsymbol{Y}^*)| \ .\end{aligned}$$

By using the fact that ℓ is bounded, and the definition of diameter $D_{\mathbb{Y}^\diamond}$, we have that

$$\begin{aligned}&|\hat{R}(\boldsymbol{h}_{\boldsymbol{\theta}^*}; X, \boldsymbol{Y}') - \hat{R}(\boldsymbol{h}_{\boldsymbol{\theta}^*}; X, \boldsymbol{Y}^*)| \\ &= \left|\frac{1}{m}\sum_{j=1}^m\sum_{c=1}^k \ell(\boldsymbol{h}_{\boldsymbol{\theta}^*}(x_j), \boldsymbol{e}_c)(y'_{jc} - y^*_{jc})\right| \\ &\leq B\frac{1}{m}\sum_{j=1}^m\sum_{c=1}^k \left|y'_{jc} - y^*_{jc}\right| \leq BD_{\mathbb{Y}^\diamond} \ .\end{aligned}$$

To wrap it up, it results that

$$\begin{aligned}R(\boldsymbol{h}_{\hat{\boldsymbol{\theta}}}) \leq &\hat{R}(\boldsymbol{h}_{\boldsymbol{\theta}^*}; X, \boldsymbol{Y}^*) + BD_{\mathbb{Y}^\diamond} + 2\hat{\mathfrak{R}}_m(\mathcal{L}; X, \boldsymbol{Y}^*) \\ &+ O\left(B\sqrt{\frac{\ln\frac{1}{\delta}}{m}}\right) \\ &\leq R(\boldsymbol{h}_{\boldsymbol{\theta}^*}) + BD_{\mathbb{Y}^\diamond} + 4\hat{\mathfrak{R}}_m(\mathcal{L}; X, \boldsymbol{Y}^*) \\ &+ O\left(B\sqrt{\frac{\ln\frac{1}{\delta}}{m}}\right) \\ &\leq R(\boldsymbol{h}_{\boldsymbol{\theta}^*}) + BD_{\mathbb{Y}^\diamond} \\ &+ \sup_{\boldsymbol{Y}\in\boldsymbol{Y}^\diamond} 4\hat{\mathfrak{R}}_m(\mathcal{L}; X, \boldsymbol{Y}) + O\left(B\sqrt{\frac{\ln\frac{1}{\delta}}{m}}\right) \ .\end{aligned}$$

Proof of Lemma 9 The proof is along the same lines of the proof of Lemma 1, but we take a union bound with respect to all the nK intervals $\triangle_{i,c,\hat{c}}$ for $i \in 1, \ldots, n$, $c \in 1, \ldots, k$, and $\hat{c} = 1, \ldots, k_i$. Moreover, as for any $j \in 1, \ldots, m$, we have that $y_{j,c}[\boldsymbol{\phi}_i(x_j)]_{\hat{c}} \leq 1$, we take $B = 1$ during the proof (as defined in Lemma 1).

B. Additional Experimental Details

We provide further information specifying the experimental setup used to generate our figures.

B.1. Weak Supervision Sources

We first build the weak supervision sources on our two datasets as follows.

Animals with Attributes. Each class is annotated with a binary vector of attributes. For each attribute, we train a binary classifier by finetuning a ResNet-18 using labeled data from the seen classes. When we consider a classification task between two unseen classes, we use as weak supervision sources the classifiers for the attributes which are different between the animals of these two unseen classes. We report the results of the 4 binary classification tasks which have the lowest majority vote accuracy. We chose these particular tasks to demonstrate the abilities of our methods on the tasks that have the least accurate weak supervision sources.

DomainNet. We sample 5 of the 25 classes of DomainNet with the largest number of datapoints. For each domain, we use 60% of the available data for those classes to fine tune a pretrained ResNet-18 network. We perform this procedure on two disjoint samples of test classes to illustrate our results on two distinct multiclass classification tasks.

In our experiments, we use the pretrained ResNet-18 from PyTorch. We finetune this ResNet-18 network following the approach described in (He et al., 2016), using cross-entropy loss.

B.2. Algorithm Hyperparameters

The subgradient method (Algorithm 1) used to train AMCL-CC and AMCL-LR uses the following hyperparameters:

AMCL-CC: We set $\delta = 0.1$, and build the constraints as in Lemma 1. We use $\varepsilon = 0.1$, and define the step size h and the number of iterations T as in Theorem 5, using $L = 2\sqrt{n}$ and diameter of Θ equal to $\sqrt{2}$.

AMCL-LR: In this case, the loss function is bounded as in Lemma 7. Since this value could be potentially very large, which in turn it would result in large intervals and number of iterations, we use the value $B = 0.1$ in the experiments. We set δ to 0.1 and build the constraints as in Lemma 1. We do not bound the set of weights Θ: in the experiments, the norm of the weights of the multinomial logistic regression model has never diverged. We run the subgradient algorithm for $T = 1000$ iterations with step size $h = 0.02$.

C. Additional Figures

C.1. Animals with Attributes

We provide the remaining figures for our experiments on the Animals with Attributes dataset. The last two binary classification tasks are bat v. rat and horse v. giraffe.

From Figure 3, we note that our methods show similar results as the figures displayed in the main body of the paper. AMCL-LR matches or outperforms all other methods on both tasks, over all ranges labeled data. AMCL-CC is within a few accuracy points of the other baselines and AMCL-LR on these tasks.

C.2. DomainNet

We provide the remaining figures for our experiments on the DomainNet dataset. We provide histograms when using the other 4 domains as the target task and also provide histograms for results on another of the samples of 5 classes. The first sample of classes as mentioned in the main body of the paper is $\{$sea turtle, vase, whale, bird, violin $\}$. The second sample is $\{$tornado, trombone, submarine, feather, zebra $\}$.

From Figures 4–8, we note that in most domains our methods perform better than or match all other approaches, namely in both samples of Clipart, Quickdraw, Painting, and the second sample of Sketch. Our methods achieve slightly lower accuracy than the best performing baseline on the Real domain and on the second sample of the Infograph domain, although they are not beaten by a single baseline in all of these tasks. We believe that the combination of our theoretical guarantees and that our methods achieve similar or sometimes better empirical performance captures the benefits of AMCL.

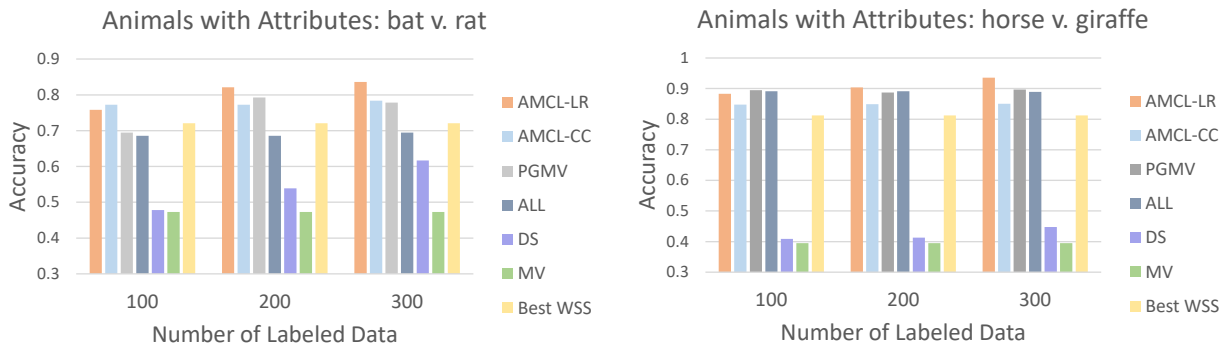

Figure 3. Experimental results on the Animals with Attributes dataset for the binary classification tasks of bat vs. rat and horse vs. giraffe as we vary the amount of labeled data. Each method uses 347 unlabeled data for bat vs. rat and 1424 unlabeled data for horse vs. giraffe.

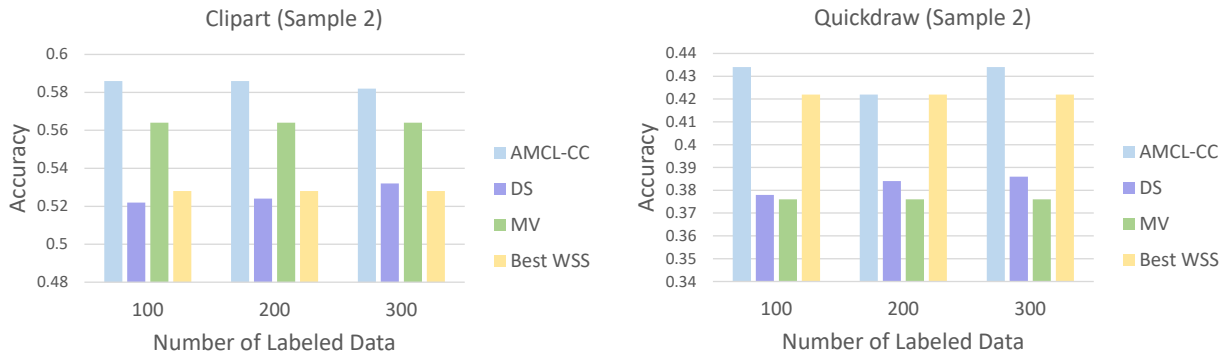

Figure 4. Experimental results on the second sample of Domain Net for the clipart and quickdraw domains as we vary the amount of labeled data. Each method uses 500 unlabeled data.

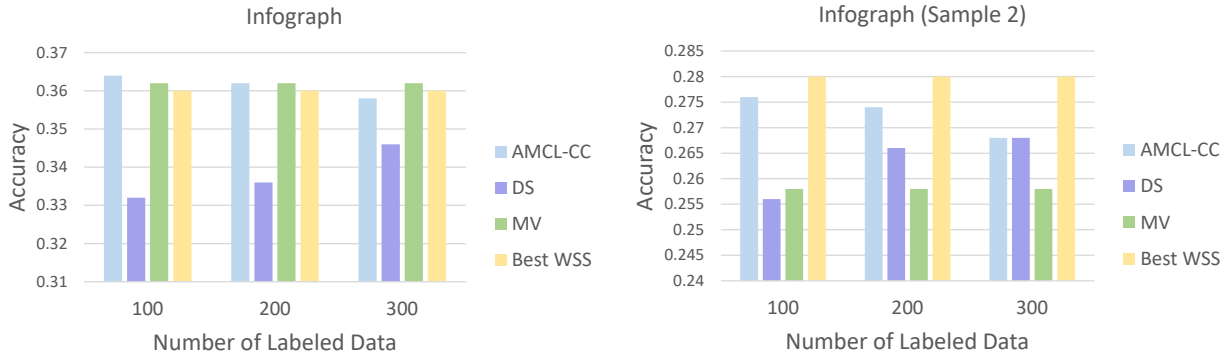

Figure 5. Experimental results on both samples of Domain Net for the Infograph domain as we vary the amount of labeled data. Each method uses 500 unlabeled data.

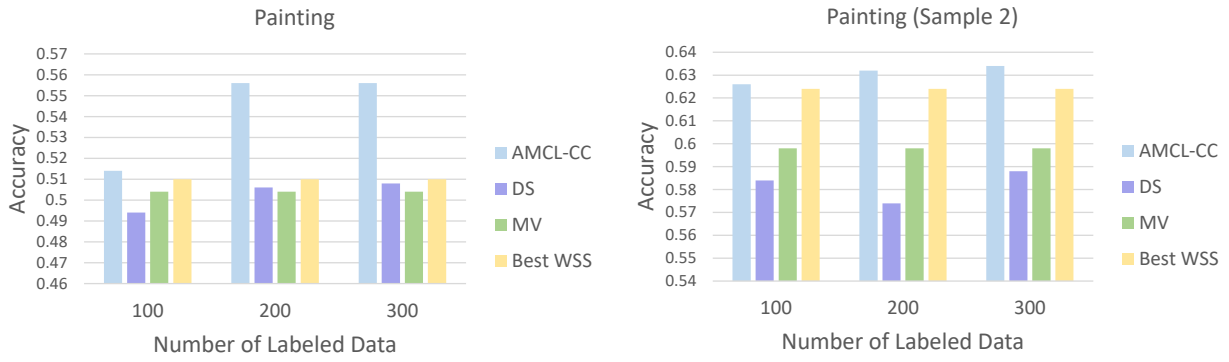

Figure 6. Experimental results on both samples of Domain Net for the Painting domain as we vary the amount of labeled data. Each method uses 500 unlabeled data.

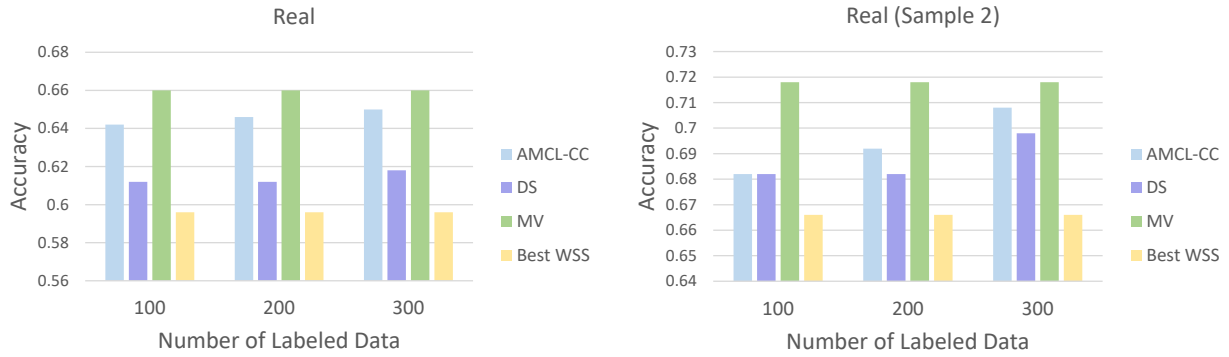

Figure 7. Experimental results on both samples of Domain Net for the Real domain as we vary the amount of labeled data. Each method uses 500 unlabeled data.

Figure 8. Experimental results on both samples of Domain Net for the Sketch domain as we vary the amount of labeled data. Each method uses 500 unlabeled data.

D. Experiments on Synthetic Data

We run synthetic experiments to show that our method is robust with respect to the addition of correlated weak supervision sources. Similar experiments have been done for ALL by Arachie & Huang (2019).

We consider a multiclass classification task over 5 classes, and 25 weak supervision sources $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_{25}$. In this classification task, each item of the domain $\mathcal{X}$ has a unique true label. Given an item $x \in \mathcal{X}$, for $i \in 1, \ldots, 10$, the weak supervision source $\boldsymbol{\phi}_i$ returns the correct label with probability $1/2$, and a random label with probability $1/2$. The output of the weak supervision source $\boldsymbol{\phi}_i$ is independent to the output of the weak supervision sources $\boldsymbol{\phi}_j$ for $j \in \{1, \ldots, 10\} \backslash \{i\}$. Therefore, the weak supervision source $\boldsymbol{\phi}_i$ is correct with probability $\frac{1}{2}(1 + \frac{1}{k})$. For $i = 11, \ldots, 25$, the weak supervision sources ϕ_i outputs the same result than the weak supervision source $\boldsymbol{\phi}_1$. Note that the weak supervision sources $\boldsymbol{\phi}_{11}, \ldots, \boldsymbol{\phi}_{25}$ do not provide any additional information with respect to the target classification task, as they add redundant constraints to the set of feasible labelings $\mathbb{Y}^\diamond$. The majority vote of the weak supervision sources $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_{25}$ is highly affected by these dependencies, and it is very likely to provide the same answer as $\boldsymbol{\phi}_1$, which is only $\frac{1}{2}(1 + \frac{1}{k})$ accurate on average. On the other hand, the majority vote of the weak supervision sources $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_{10}$ would improve upon the individual accuracy of the weak supervision sources, as their output is independent.

We use 500 unlabeled examples, run experiments varying the amount of labeled data, and show that our method AMCL-CC is robust against those dependencies. For the sake of these experiments, as we want to use very small amount of labeled data, we set $\gamma = 0$ when building the constraints for $\mathbb{Y}^\diamond$ as in Lemma 1. The experimental results are reported in Table D. The table shows that AMCL-CC is robust with respect to dependencies among weak supervision sources, whereas majority vote is greatly affected by them. In fact, in this case the majority vote does not improve upon the individual accuracy of the weak supervision sources, which is on average $\frac{1}{2}(1 + \frac{1}{k}) = \frac{3}{5}$.

Table 1. We report the experimental results on the synthetic dataset. We report the accuracy obtained by our method AMCL-CC and the majority vote, when varying the amount of labeled examples (we report the average accuracy over 3 distinct runs).

Labeled Examples	AMCL-CC	Majority Vote
100	0.902	0.595
50	0.828	0.602
25	0.819	0.598