Active Learning on Attributed Graphs via Graph Cognizant Logistic Regression and Preemptive Query Generation
- Supplementary Material -

Florence Regol[1] Soumyasundar Pal[1] Yingxue Zhang[2] Mark Coates[1]

1. Active Learning proof

Let $\{\mathbf{x}_i \in \mathbb{R}^d : 1 \leq i \leq N\}$ be N feature vectors and denote the indices of the vectors by $\mathcal{D} = \{1, 2, ...N\}$. Associated with the feature $\mathbf{x}_i$ are two binary class labels $\mathbf{y}_{i,1}$ and $\mathbf{y}_{i,2}$. These values differ for exactly $m < N$ of the feature vectors, with indices $\{i_k\}, k = 1, \ldots, m$. We learn two logistic regression models with the same regularization parameter λ as follows:

$$\begin{aligned}\mathbf{w}_1 &= \arg\min_{\mathbf{w}\in\mathbb{R}^d} \sum_{i\in\mathcal{D}} \ell(\mathbf{x}_i, \mathbf{y}_{i,1}; \mathbf{w}) + \frac{\lambda}{2}||\mathbf{w}||_2^2 \,, \\ \mathbf{w}_2 &= \arg\min_{\mathbf{w}\in\mathbb{R}^d} \sum_{i\in\mathcal{D}} \ell(\mathbf{x}_i, \mathbf{y}_{i,2}; \mathbf{w}) + \frac{\lambda}{2}||\mathbf{w}||_2^2 \,. \end{aligned} \tag{1}$$

Here $\ell(\mathbf{x}_i, \mathbf{y}_i; \mathbf{w}) = -\Big(\mathbf{y}_i \log \sigma(\mathbf{w}^\top \mathbf{x}_i) + \big(1 - \mathbf{y}_i\big) \log \big(1 - \sigma(\mathbf{w}^\top \mathbf{x}_i)\big)\Big)$ is the cross entropy for an individual training example and $\sigma(x) = \dfrac{1}{1 + e^{-x}}$ is the sigmoid function.

Theorem 1 of (Sivan et al., 2019) provides a bound for the difference between the solutions of optimization problems of the form in (1). For the special case of L2-regularized binary logistic regression with m different labels, we have $||\mathbf{w}_1 - \mathbf{w}_2|| \leq 2||r||$, where

$$\begin{aligned} r &\triangleq \frac{1}{2\lambda} \sum_{k=1}^{m} \Bigg(\nabla_w \ell(\mathbf{x}_{i_k}, \mathbf{y}_{i_k,1}; \mathbf{w})|_{\mathbf{w}=\mathbf{w}_1} - \nabla_w \ell(\mathbf{x}_{i_k}, \mathbf{y}_{i_k,2}; \mathbf{w})|_{\mathbf{w}=\mathbf{w}_1} \Bigg) \,, \\ &= \frac{1}{2\lambda} \sum_{k=1}^{m} \Bigg(-\big(\mathbf{y}_{i_k,1} - \sigma(\mathbf{w}_1^\top \mathbf{x}_{i_k})\big)\mathbf{x}_{i_k} + \big(\mathbf{y}_{i_k,2} - \sigma(\mathbf{w}_1^\top \mathbf{x}_{i_k})\big)\mathbf{x}_{i_k} \Bigg) \,, \\ &= \frac{1}{2\lambda} \sum_{k=1}^{m} \big(\mathbf{y}_{i_k,2} - \mathbf{y}_{i_k,1}\big)\mathbf{x}_{i_k} \,. \end{aligned} \tag{2}$$

Lemma 1 from (Sivan et al., 2019) provides upper and lower bounds for the prediction associated with a sample $\mathbf{x}_j$ when using the weights $\mathbf{w}_2$ derived from labels $\{\mathbf{y}_{i,2}\}$:

$$\mathbf{w}_1^\top \mathbf{x}_j - r^\top \mathbf{x}_j - ||r||.||\mathbf{x}_j|| \leq \mathbf{w}_2^\top \mathbf{x}_j \leq \mathbf{w}_1^\top \mathbf{x}_j - r^\top \mathbf{x}_j + ||r||.||\mathbf{x}_j|| \,. \tag{3}$$

Based on these bounds we can specify the following proposition:

Proposition 1. *For weights $\mathbf{w}_1$ and $\mathbf{w}_2$ derived by L2-penalized logistic regression to two datasets with common feature vectors but label sets differing for m vectors indexed by $\{i_k\}, k = 1, \ldots, m$, define $\eta \triangleq \frac{1}{2\lambda}\sum_{k=1}^{m} ||\mathbf{x}_{i_k}||$ and $b_{\pm\eta}(\mathbf{w}_2, j) \triangleq$*

[1] McGill University, Montréal, Canada [2] Huawei, Montréal, Canada. Correspondence to: Florence Regol <florence.robert-regol@mail.mcgill.ca>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020. Copyright 2020 by the author(s).

$\sigma(\mathbf{w}_2^\top \mathbf{x}_j) - \sigma(\mathbf{w}_2^\top \mathbf{x}_j \pm 2\eta||\mathbf{x}_j||)$. *For any* $\mathbf{x}_j$*, for* $j \in \{1, \dots, N\}$*:*

$$|\sigma(\mathbf{w}_1^\top \mathbf{x}_j) - \sigma(\mathbf{w}_2^\top \mathbf{x}_j)| \leq \max(|b_\eta(\mathbf{w}_2, j)|, |b_{-\eta}(\mathbf{w}_2, j)|)\,. \tag{4}$$

Proof. We first note that $\big(\mathbf{y}_{i_k,2} - \mathbf{y}_{i_k,1}\big) \in \{-1, 1\}$ implies from the definition of r in (2) that $||r|| \leq \dfrac{1}{2\lambda}\displaystyle\sum_{k=1}^{m}||\mathbf{x}_{i_k}|| = \eta$. Applying the Cauchy Schwarz inequality to (3), we have:

$$\mathbf{w}_2^\top \mathbf{x}_j - 2||r||.||\mathbf{x}_j|| \leq \mathbf{w}_1^\top \mathbf{x}_j \leq \mathbf{w}_2^\top \mathbf{x}_j + 2||r||.||\mathbf{x}_j||\,, \tag{5}$$

and hence

$$\mathbf{w}_2^\top \mathbf{x}_j - 2\eta||\mathbf{x}_j|| \leq \mathbf{w}_1^\top \mathbf{x}_j \leq \mathbf{w}_2^\top \mathbf{x}_j + 2\eta||\mathbf{x}_j||\,. \tag{6}$$

Since $\sigma(\cdot)$ is monotonically increasing, we have $\sigma(\mathbf{w}_1^\top \mathbf{x}_j - 2\eta||\mathbf{x}_j||) \leq \sigma(\mathbf{w}_2^\top \mathbf{x}_j) \leq \sigma(\mathbf{w}_1^\top \mathbf{x}_j + 2\eta||\mathbf{x}_j||)$ and (4) follows. □

1.1. Useful results

The proof of the main result bounding the difference in risks relies on the following two lemmas:

Lemma 1.1. *Let* X *and* Y *be two random variables taking values in* $[a, b]$ *and* $[c, d]$ *respectively. Then* $|\mathbf{E}_X[X] - \mathbf{E}_Y[Y]| \leq \max\Big(|a - d|, |b - c|\Big)$.

Proof. We note

$$a - d \leq \mathbf{E}_X[X] - \mathbf{E}_Y[Y] \leq b - c\,,$$

from which the result follows. □

Lemma 1.2. *If* $0 \leq p_1, p_2 \leq 1$*, then* $|\min(p_1, 1 - p_1) - \min(p_2, 1 - p_2)| \leq |p_1 - p_2|$.

Proof. Let $q = \min(p_1, 1 - p_1) - \min(p_2, 1 - p_2)$.

$$\begin{aligned}
&\text{case 1}: p_1 < 0.5, \quad p_2 < 0.5\\
&\qquad q = p_1 - p_2\\
&\text{case 2}: p_1 \geq 0.5, \quad p_2 \geq 0.5\\
&\qquad q = (1 - p_1) - (1 - p_2) = p_2 - p_1\\
&\text{case 3}: p_1 < 0.5, \quad p_2 \geq 0.5\\
&\qquad p_1 - p_2 \leq q = p_1 - (1 - p_2) \leq (1 - p_1) - (1 - p_2) = p_2 - p_1\\
&\text{case 4}: p_1 \geq 0.5, \quad p_2 < 0.5\\
&\qquad p_2 - p_1 = (1 - p_1) - (1 - p_2) \leq q = (1 - p_1) - p_2 \leq p_1 - p_2
\end{aligned}$$

So, we have $|q| \leq |p_1 - p_2|$. □

1.2. Risk bound for binary classification

We now recall the definition of the risk of querying node q given a current label set $\mathbf{Y}_{\mathcal{L}_t}$ for binary classification:

$$R^{+q}_{|\mathbf{Y}_{\mathcal{L}_t}} \triangleq \tfrac{1}{|\mathcal{U}_t^{-q}|}\sum_{k \in \{0,1\}}\sum_{i \in \mathcal{U}_t^{-q}} \varphi^{+q}_{i,k,\mathbf{Y}_{\mathcal{L}_t}}\, p(\mathbf{y}_q = k|\mathbf{Y}_{\mathcal{L}_t})\,, \tag{7}$$

where $\varphi^{+q}_{i,k,\mathbf{Y}_{\mathcal{L}_t}} \triangleq \Big(1 - \max\limits_{k' \in \{0,1\}} p(\mathbf{y}_i = k'|\mathbf{Y}_{\mathcal{L}_t}, \mathbf{y}_q = k)\Big) = \min\limits_{k' \in \{0,1\}} p(\mathbf{y}_i = k'|\mathbf{Y}_{\mathcal{L}_t}, \mathbf{y}_q = k)$. We are interested in bounding the difference in the risks that can arise when we use two label sets $\mathbf{Y}_{\mathcal{L}_t}$ and $\mathbf{Y}'_{\mathcal{L}_t}$ that differ by only one label, associated with the node index q^*_{t-1}.

We now state the main result:

Theorem 1.3. *The risk error arising from applying binary L2-regularized logistic regression with regularization parameter* λ *to two labelled datasets* $\mathbf{Y}_{\mathcal{L}_t}$ *and* $\mathbf{Y}'_{\mathcal{L}_t}$ *that differ by one label, associated with the node* q^*_{t-1}*, is bounded as:*

$$|R^{+q}_{|\mathbf{Y}_{\mathcal{L}_t}} - R^{+q}_{|\mathbf{Y}'_{\mathcal{L}_t}}| \leq \frac{1}{|\mathcal{U}_t^{-q}|} \sum_{i \in \mathcal{U}_t^{-q}} \tilde{b}(\eta_q, \mathbf{w}_2, i)\,,$$

where $\mathbf{w}_{2,k}$ *is the weight vector learned using* $\{\mathbf{Y}'_{\mathcal{L}_t} \cup \{\mathbf{y}_q = k\}\}$*,* $\eta_q \triangleq \frac{1}{2\lambda}\Big(||\mathbf{x}_q|| + ||\mathbf{x}_{q^*_{t-1}}||\Big)$*, and*

$$\tilde{b}(\eta_q, \mathbf{w}_2, i) = \max_{k \in \{0,1\}} \max\left(\left|\sigma(\mathbf{w}_{2,k}^\top \mathbf{x}_i) - \sigma(\mathbf{w}_{2,k}^\top \mathbf{x}_i + 2\eta_q||\mathbf{x}_i||)\right|, \left|\sigma(\mathbf{w}_{2,k}^\top \mathbf{x}_i) - \sigma(\mathbf{w}_{2,k}^\top \mathbf{x}_i - 2\eta_q||\mathbf{x}_i||)\right|\right)\,. \tag{8}$$

Proof. We define the random variable $\varphi^{+q}_{i,\mathbf{Y}_{\mathcal{L}_t}}$ which takes value $\varphi^{+q}_{i,k,\mathbf{Y}_{\mathcal{L}_t}}$ with probability $p(\mathbf{y}_q = k|\mathbf{Y}_{\mathcal{L}_t})$ for $k \in \{0,1\}$. Analogously, let $\varphi^{+q}_{i,\mathbf{Y}'_{\mathcal{L}_t}}$ take value $\varphi^{+q}_{i,k,\mathbf{Y}'_{\mathcal{L}_t}}$ with probability $p(q = k|\mathbf{Y}'_{\mathcal{L}_t})$ for $k \in \{0,1\}$. The difference in risk is then:

$$R^{+q}_{|\mathbf{Y}_{\mathcal{L}_t}} - R^{+q}_{|\mathbf{Y}'_{\mathcal{L}_t}} = \tfrac{1}{|\mathcal{U}_t^{-q}|} \sum_{i \in \mathcal{U}_t^{-q}} E_{\mathbf{y}_q}[\varphi^{+q}_{i,\mathbf{Y}_{\mathcal{L}_t}}] - E_{\mathbf{y}_q}[\varphi^{+q}_{i,\mathbf{Y}'_{\mathcal{L}_t}}]\,, \tag{9}$$

where $E_{\mathbf{y}_q}$ denotes expectation over $\mathbf{y}_q$ conditioned on the corresponding observed label sets, either $\mathbf{Y}_{\mathcal{L}_t}$ or $\mathbf{Y}'_{\mathcal{L}_t}$. For query node q, for each $k_1, k_2 \in \{0,1\}$, we learn weights $\mathbf{w}_{1,k_1}$ using $\{\mathbf{Y}_{\mathcal{L}_t} \cup \{\mathbf{y}_q = k_1\}\}$ and weights $\mathbf{w}_{2,k_2}$ using $\{\mathbf{Y}'_{\mathcal{L}_t} \cup \{\mathbf{y}_q = k_2\}\}$. For each $i \in \mathcal{U}_t^{-q}$, we have:

$$\begin{aligned} |\varphi^{+q}_{i,k_1,\mathbf{Y}_{\mathcal{L}_t}} - \varphi^{+q}_{i,k_2,\mathbf{Y}'_{\mathcal{L}_t}}| &\leq |\sigma(\mathbf{w}_{1,k_1}^\top \mathbf{x}_i) - \sigma(\mathbf{w}_{2,k_2}^\top \mathbf{x}_i)|\,,\ \text{using Lemma (1.2)}\,, \\ &\leq \max(|\sigma(\mathbf{w}_{2,k_2}^\top \mathbf{x}_i) - \sigma(\mathbf{w}_{2,k_2}^\top \mathbf{x}_i - 2\eta_q||\mathbf{x}_i||)|, |\sigma(\mathbf{w}_{2,k_2}^\top \mathbf{x}_i) - \sigma(\mathbf{w}_{2,k_2}^\top \mathbf{x}_i + 2\eta_q||\mathbf{x}_i||)|)\,, \\ &\leq \max_{k \in \{0,1\}} \max(|\sigma(\mathbf{w}_{2,k}^\top \mathbf{x}_i) - \sigma(\mathbf{w}_{2,k}^\top \mathbf{x}_i - 2\eta_q||\mathbf{x}_i||)|, |\sigma(\mathbf{w}_{2,k}^\top \mathbf{x}_i) - \sigma(\mathbf{w}_{2,k}^\top \mathbf{x}_i + 2\eta_q||\mathbf{x}_i||)|)\,. \end{aligned} \tag{10}$$

Here the second inequality follows from Proposition (1), observing that the labels can differ for nodes q^*_{t-1} and q.

Now, the difference in risk is bounded as

$$\begin{aligned} &|R^{+q}_{|\mathbf{Y}_{\mathcal{L}_t}} - R^{+q}_{|\mathbf{Y}'_{\mathcal{L}_t}}| \leq \frac{1}{|\mathcal{U}_t^{-q}|} \sum_{i \in \mathcal{U}_t^{-q}} |E_{\mathbf{y}_q}[\varphi^{+q}_{i,\mathbf{Y}_{\mathcal{L}_t}}] - E_{\mathbf{y}_q}[\varphi^{+q}_{i,\mathbf{Y}'_{\mathcal{L}_t}}]|\,, \\ &\leq \frac{1}{|\mathcal{U}_t^{-q}|} \sum_{i \in \mathcal{U}_t^{-q}} \max(|\max_{k \in \{0,1\}} \varphi^{+q}_{i,k,\mathbf{Y}_{\mathcal{L}_t}} - \min_{k \in \{0,1\}} \varphi^{+q}_{i,k,\mathbf{Y}'_{\mathcal{L}_t}}|, \quad |\min_{k \in \{0,1\}} \varphi^{+q}_{i,k,\mathbf{Y}_{\mathcal{L}_t}} - \max_{k \in \{0,1\}} \varphi^{+q}_{i,k,\mathbf{Y}'_{\mathcal{L}_t}}|)\,,\ \text{using Lemma (1.1)}\,, \\ &\leq \frac{1}{|\mathcal{U}_t^{-q}|} \sum_{i \in \mathcal{U}_t^{-q}} \max_{k \in \{0,1\}} \max(|\sigma(\mathbf{w}_{2,k}^\top \mathbf{x}_i) - \sigma(\mathbf{w}_{2,k}^\top \mathbf{x}_i - 2\eta_q||\mathbf{x}_i||)|, |\sigma(\mathbf{w}_{2,k}^\top \mathbf{x}_i) - \sigma(\mathbf{w}_{2,k}^\top \mathbf{x}_i + 2\eta_q||\mathbf{x}_i||)|)\,,\ \text{using (10)}\,, \\ &= \frac{1}{|\mathcal{U}_t^{-q}|} \sum_{i \in \mathcal{U}_t^{-q}} \tilde{b}(\eta_q, \mathbf{w}_2, i) \end{aligned} \tag{11}$$

□

1.3. Risk bound for multiclass classification

We now consider the case where we perform multi-class classification using one-versus-all binary logistic regression for each class and then normalizing by the sum of the output sigmoids. As before, we focus on the case where two labels can be different for the proposed query node q and the previous query node q^*_{t-1}.

For a given label set Y, we learn weights $\mathbf{w}^{(k)}$ for each class $k \in \{1,\ldots,K\}$ using L2-regularized binary one-vs-all logistic regression. The output prior to normalization for a given feature vector $\mathbf{x}_i$ is then $\sigma(\mathbf{w}^{(k)\top}\mathbf{x}_i)$. We then normalize by dividing by $C_i = \sum_{k=1}^K \sigma(\mathbf{w}^{(k)\top}\mathbf{x}_i)$ to obtain a probability vector v.

Define

$$\beta(\mathbf{w}_2, \eta, i) \triangleq \max_k \beta(\mathbf{w}_2, k, \eta, i), \tag{12}$$

for

$$\beta(\mathbf{w}_2, k, \eta, i) \triangleq \max\left(\left| \frac{\sigma(\mathbf{w}_2^{(k)\top}\mathbf{x}_i)}{C_{2,i}} - \frac{\sigma(\mathbf{w}_2^{(k)\top}\mathbf{x}_i) - \rho(\mathbf{w}_2, \eta, i)}{C_{2,i} + 4\rho(\mathbf{w}_2, \eta, i)} \right|, \left| \frac{\sigma(\mathbf{w}_2^{(k)\top}\mathbf{x}_i)}{C_{2,i}} - \frac{\sigma(\mathbf{w}_2^{(k)\top}\mathbf{x}_i) + \rho(\mathbf{w}_2, \eta, i)}{C_{2,i} - 4\rho(\mathbf{w}_2, \eta, i)} \right| \right), \tag{13}$$

where $\rho(\mathbf{w}_2, \eta, i) \triangleq \max_k \max(|b_\eta(\mathbf{w}_2^{(k)}, i)|, |b_{-\eta}(\mathbf{w}_2^{(k)}, i)|)$ for $b_{\pm\eta}(\mathbf{w}_2, i) \triangleq \sigma(\mathbf{w}_2^\top \mathbf{x}_i) - \sigma(\mathbf{w}_2^\top \mathbf{x}_i \pm 2\eta||\mathbf{x}_i||)$.

Proposition 2. *For two label sets Y and $\mathbf{Y}'$ that differ in exactly two labels, $\mathbf{y}_{j_1} \neq \mathbf{y}'_{j_1}$ and $\mathbf{y}_{j_2} \neq \mathbf{y}'_{j_2}$, we perform multiclass regression for classes $k \in \{1, \ldots, K\}$ by conducting L2-regularized one-vs-all binary logistic regression for each class and then normalizing the output values to sum to one. Let $\mathbf{w}_1^{(k)}$ denote the weights learned for class k using label set Y and $\mathbf{w}_2^{(k)}$ denote the corresponding weights learned using label set $\mathbf{Y}'$. Let $p_{1,k,i} = \sigma(\mathbf{w}_1^{(k)\top}\mathbf{x}_i)/C_{1,i}$ be the output probability associated with class k for feature vector $\mathbf{x}_i$ using label set Y, where $C_{1,i} = \sum_{k=1}^K \sigma(\mathbf{w}_1^{(k)\top}\mathbf{x}_i)$. Let $p_{2,k,i}$ and $C_{2,i}$ be the corresponding entities derived using label set $\mathbf{Y}'$.*

For any feature vector $\mathbf{x}_i$, $|p_{1,k,i} - p_{2,k,i}| \leq \beta(\mathbf{w}_2, \eta_s, i)$ for each class k where $\eta_s \triangleq \frac{1}{2\lambda}(||\mathbf{x}_{j_1}|| + ||\mathbf{x}_{j_2}||)$.

Proof. Two different labels can affect the weight vectors $\mathbf{w}^{(k)}$ for at most four classes, because there is no change in the one-versus-all binary labels for class k unless at least one of $\mathbf{y}_{j_1}$, $\mathbf{y}_{j_2}$, $\mathbf{y}'_{j_1}$ or $\mathbf{y}'_{j_2}$ equals k. For any class k, at most two labels can change for the binary classification. For the one-vs-all binary classifier for class k, we learn a weight vector $\mathbf{w}_1^{(k)}$ using Y and $\mathbf{w}_2^{(k)}$ using $\mathbf{Y}'$. Then from Proposition 1, we have (for the case where both labels change):

$$|\sigma(\mathbf{w}_1^{(k)\top}\mathbf{x}_i) - \sigma(\mathbf{w}_2^{(k)\top}\mathbf{x}_i)| \leq \max(|b_{\eta_s}(\mathbf{w}_2^{(k)}, i)|, |b_{-\eta_s}(\mathbf{w}_2^{(k)}, j)|)\,. \tag{14}$$

Define $\rho(\mathbf{w}_2, \eta, i) \triangleq \max_k \max(|b_\eta(\mathbf{w}_2^{(k)}, i)|, |b_{-\eta}(\mathbf{w}_2^{(k)}, i)|)$. Then for any class k,

$$\sigma(\mathbf{w}_2^{(k)\top}\mathbf{x}_i) - \rho(\mathbf{w}_2, \eta_s, i) \leq \sigma(\mathbf{w}_1^{(k)\top}\mathbf{x}_i) \leq \sigma(\mathbf{w}_2^{(k)\top}\mathbf{x}_i) + \rho(\mathbf{w}_2, \eta_s, i)\,. \tag{15}$$

We can thus bound the difference in the normalization terms $C_{1,i} = \sum_{k=1}^K \sigma(\mathbf{w}_1^{(k)\top}\mathbf{x}_i)$ and $C_{2,i} = \sum_{k=1}^K \sigma(\mathbf{w}_2^{(k)\top}\mathbf{x}_i)$. Since at most four classes can be affected by the change of two labels, we have:

$$C_{2,i} - 4\rho(\mathbf{w}_2, \eta_s, i) \leq C_{1,i} \leq C_{2,i} + 4\rho(\mathbf{w}_2, \eta_s, i)\,. \tag{16}$$

We can now bound the difference in the probabilities $p_{1,k,i} = \sigma(\mathbf{w}_1^{(k)\top}\mathbf{x}_i)/C_{1,i}$ and $p_{2,k,i} = \sigma(\mathbf{w}_2^{(k)\top}\mathbf{x}_i)/C_{2,i}$ as follows:

$$\begin{aligned}|p_{1,k,i} - p_{2,k,i}| &= |\sigma(\mathbf{w}_2^{(k)\top}\mathbf{x}_i)/C_{2,i} - \sigma(\mathbf{w}_1^{(k)\top}\mathbf{x}_i)/C_{1,i}| \\ &\leq \max\left(\left| \frac{\sigma(\mathbf{w}_2^{(k)\top}\mathbf{x}_i)}{C_{2,i}} - \frac{\sigma(\mathbf{w}_2^{(k)\top}\mathbf{x}_i) - \rho(\mathbf{w}_2, \eta_s, i)}{C_{2,i} + 4\rho(\mathbf{w}_2, \eta_s, i)} \right|, \left| \frac{\sigma(\mathbf{w}_2^{(k)\top}\mathbf{x}_i)}{C_{2,i}} - \frac{\sigma(\mathbf{w}_2^{(k)\top}\mathbf{x}_i) + \rho(\mathbf{w}_2, \eta_s, i)}{C_{2,i} - 4\rho(\mathbf{w}_2, \eta_s, i)} \right| \right)\end{aligned} \tag{17}$$

Taking the maximum of the right hand side of (17) over k leads to $|p_{1,k,i} - p_{2,k,i}| < \beta(\mathbf{w}_2, \eta_s, i)$ for all k. □

Recalling the definition of the risk of querying node q given a current label set $\mathbf{Y}_{\mathcal{L}_t}$ for multiclass classification:

$$R^{+q}_{|\mathbf{Y}_{\mathcal{L}_t}} \triangleq \tfrac{1}{|\mathcal{U}_t^{-q}|} \sum_{k \in \{1,\ldots,K\}} \sum_{i \in \mathcal{U}_t^{-q}} \varphi^{+q}_{i,k,\mathbf{Y}_{\mathcal{L}_t}} p(\mathbf{y}_q = k|\mathbf{Y}_{\mathcal{L}_t})\,, \tag{18}$$

where $\varphi^{+q}_{i,k,\mathbf{Y}_{\mathcal{L}_t}} \triangleq \left(1 - \max_{k' \in \{1,\ldots,K\}} p(\mathbf{y}_i = k'|\mathbf{Y}_{\mathcal{L}_t}, \mathbf{y}_q = k)\right)$. As for the binary case, the difference in risk is:

$$R^{+q}_{|\mathbf{Y}_{\mathcal{L}_t}} - R^{+q}_{|\mathbf{Y}'_{\mathcal{L}_t}} = \tfrac{1}{|\mathcal{U}_t^{-q}|} \sum_{i \in \mathcal{U}_t^{-q}} E_{\mathbf{y}_q}[\varphi^{+q}_{i,\mathbf{Y}_{\mathcal{L}_t}}] - E_{\mathbf{y}_q}[\varphi^{+q}_{i,\mathbf{Y}'_{\mathcal{L}_t}}]\,, \tag{19}$$

where $E_{\mathbf{y}_q}$ denotes expectation over $\mathbf{y}_q$ conditioned on the observed label set, either $\mathbf{Y}_{\mathcal{L}_t}$ or $\mathbf{Y}_{\mathcal{L}_t}$. The random variable $\varphi^{+q}_{i,\mathbf{Y}_{\mathcal{L}_t}}$ now takes on value $\varphi^{+q}_{i,k,\mathbf{Y}_{\mathcal{L}_t}}$ with probability $p(\mathbf{y}_q = k|\mathbf{Y}_{\mathcal{L}_t})$ for $k \in \{1, \ldots, K\}$.

Theorem 1.4. *Consider multiclass regression performed via repeated one-vs-all L2-regularized logistic regression with regularization parameter λ to two labelled datasets $\mathbf{Y}_{\mathcal{L}_t}$ and $\mathbf{Y}'_{\mathcal{L}_t}$ that differ by one label, associated with the node q^*_{t-1}. Let $\mathbf{w}^{(k')}_{2,k}$ be the weight vector learned for class k' using label data $\{\mathbf{Y}'_{\mathcal{L}_t} \cup \{\mathbf{y}_q = k\}$. The risk error arising from using $\mathbf{Y}'_{\mathcal{L}_t}$ instead of $\mathbf{Y}_{\mathcal{L}_t}$ is bounded as:*

$$|R^{+q}_{|\mathbf{Y}_{\mathcal{L}_t}} - R^{+q}_{|\mathbf{Y}'_{\mathcal{L}_t}}| \leq \frac{1}{|\mathcal{U}_t^{-q}|} \sum_{i \in \mathcal{U}_t^{-q}} \tilde{\beta}(\mathbf{w}_2, \eta_q, i)\,. \tag{20}$$

Here $\tilde{\beta}(\mathbf{w}_2, \eta_q, i) = \max_k \beta(\mathbf{w}_{2,k}, \eta_q, i)$ for $\beta(\mathbf{w}_{2,k}, \eta_q, i)$ as defined in (12) and $\eta_q \triangleq \frac{1}{2\lambda}\big(||\mathbf{x}_q|| + ||\mathbf{x}_{q^*_{t-1}}||\big)$.

Proof. For query node q, for each $k_1, k_2 \in \{1, \dots, K\}$, we learn weights $\mathbf{w}^{(k)}_{1,k_1}$ using $\{\mathbf{Y}_{\mathcal{L}_t} \cup \{\mathbf{y}_q = k_1\}\}$ and $\mathbf{w}^{(k)}_{2,k_2}$ using $\{\mathbf{Y}'_{\mathcal{L}_t} \cup \{\mathbf{y}_q = k_2\}\}$, for each candidate class k. For each $i \in \mathcal{U}_t^{-q}$, we have:

$$\begin{aligned}
|\varphi^{+q}_{i,k_1,\mathbf{Y}_{\mathcal{L}_t}} - \varphi^{+q}_{i,k_2,\mathbf{Y}'_{\mathcal{L}_t}}| &= \left| \max_{k' \in \{1,\dots,K\}} p(\mathbf{y}_i = k' | \mathbf{Y}'_{\mathcal{L}_t}, \mathbf{y}_q = k_2) - \max_{k' \in \{1,\dots,K\}} p(\mathbf{y}_i = k' | \mathbf{Y}_{\mathcal{L}_t}, \mathbf{y}_q = k_1) \right| , \\
&\leq \max \left(\max_{k' \in \{1,\dots,K\}} |p(\mathbf{y}_i = k' | \mathbf{Y}'_{\mathcal{L}_t}, \mathbf{y}_q = k_2) - p(\mathbf{y}_i = k' | \mathbf{Y}_{\mathcal{L}_t}, \mathbf{y}_q = k_1)| \right) , \\
&\leq \beta(\mathbf{w}_{2,k_2}, \eta_q, i)\,, \\
&\leq \max_{k_2} \beta(\mathbf{w}_{2,k_2}, \eta_q, i) \triangleq \tilde{\beta}(\mathbf{w}_2, \eta_q, i)\,.
\end{aligned} \tag{21}$$

where the last line follows from Proposition 2.

Now, the difference in risk is bounded as

$$\begin{aligned}
&|R^{+q}_{|\mathbf{Y}_{\mathcal{L}_t}} - R^{+q}_{|\mathbf{Y}'_{\mathcal{L}_t}}| \leq \frac{1}{|\mathcal{U}_t^{-q}|} \sum_{i \in \mathcal{U}_t^{-q}} |E_{\mathbf{y}_q}[\varphi^{+q}_{i,\mathbf{Y}_{\mathcal{L}_t}}] - E_{\mathbf{y}_q}[\varphi^{+q}_{i,\mathbf{Y}'_{\mathcal{L}_t}}]|\,, \\
&\leq \frac{1}{|\mathcal{U}_t^{-q}|} \sum_{i \in \mathcal{U}_t^{-q}} \max(|\max_k \varphi^{+q}_{i,k,\mathbf{Y}_{\mathcal{L}_t}} - \min_k \varphi^{+q}_{i,k,\mathbf{Y}_{\mathcal{L}_t}}|, \quad |\min_k \varphi^{+q}_{i,k,\mathbf{Y}_{\mathcal{L}_t}} - \max_k \varphi^{+q}_{i,k,\mathbf{Y}'_{\mathcal{L}_t}}|)\,, \text{ using Lemma (1.1)}\,, \\
&\leq \frac{1}{|\mathcal{U}_t^{-q}|} \sum_{i \in \mathcal{U}_t^{-q}} \tilde{\beta}(\mathbf{w}_2, \eta_q, i)\,, \text{ using (21)}\,.
\end{aligned} \tag{22}$$

□

2. Computation of model evidence from BMRF

In this section, we provide more details for the computation of the model evidence of the BMRF model $\lambda_{TSA,\mathbf{Y}_{\mathcal{L}}}$. We consider an arbitrary ordering of the nodes in the labelled set and obtain the joint probability of interest $p(\mathbf{Y}_{\mathcal{L}})$ with a chain rule of conditionals as follow :

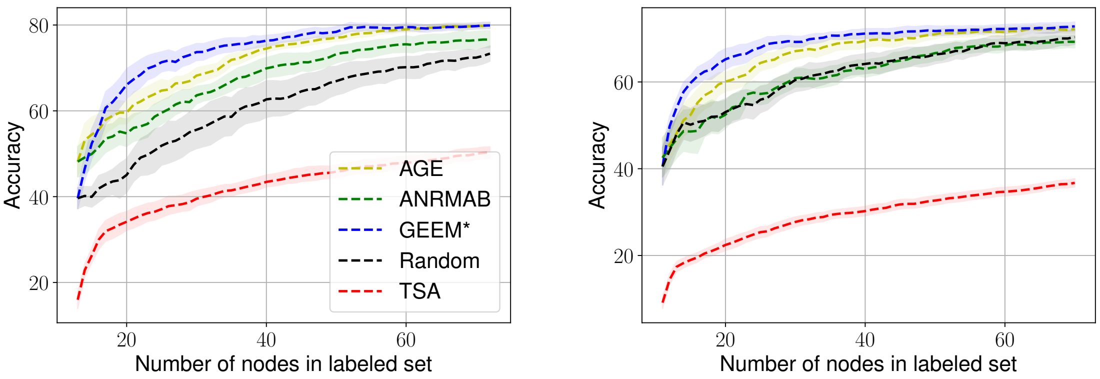
$$p(\mathbf{Y}_{\mathcal{L}}) = p(\mathbf{y}_1)p(\mathbf{y}_2|\mathbf{y}_1)...p(\mathbf{y}_{|\mathcal{L}|}|\mathbf{y}_1, \mathbf{y}_2, ...\mathbf{y}_{|\mathcal{L}|-1})\,. \tag{23}$$

The conditional distributions are evaluated with the approximation provided in (Jun & Nowak, 2016). The model evidence can be updated at each active learning iteration by only computing the conditional probability of the newly added point, which is already computed as part of the TSA active learning algorithm. As a result, the added time complexity results from the computation of $p(\mathbf{Y}_{\mathcal{L}_0})$ at the beginning of the algorithm. The marginal probability of the first node label $p(\mathbf{y}_1)$ is set uniformly for all classes.

3. Additional Experiments

In this section we report the results for Experiment 1 but include the best-performing label propagation technique (TSA (Jun & Nowak, 2016)). This is the setting where the initial training set is larger than only one node.

These results are provided to illustrate that label propagation methods are not competitive with the GCN-based methods when a larger initial training set is available. Figure 2 shows the results for the Cora and Citeseer datasets.



(a) Cora

(b) Citeseer

Figure 1. Experiment 1 with added label-propagation baseline TSA (Jun & Nowak, 2016).

4. SGCN extension : Simplified Chebynet for classification

In the main paper, we employed the SGC, a form of graph-cognizant logistic regression that can be derived as a linear version of the GCN (Kipf & Welling, 2017), as described in (Wu et al., 2019). For some graph datasets, the GCN is not a good model, because it relies strongly on immediate neighbours being similar to the centre node. In this section we derive an alternate graph-cognizant logistic regression architecture that is equivalent to a linear version of the Chebynet (Defferrard et al., 2016).

4.1. Methodology

Following the same methodology as in (Wu et al., 2019), we can also derive a linear version of the Chebynet (Defferrard et al., 2016). This graph convolutional neural network incorporates a more general version of graph convolution. This can be useful for datasets that have a more complex structure which cannot be captured by the approximate first order GCN model. A 2-layer Chebynet has the form:

$$\sigma\left(\sum_{w=0}^{W-1}\mathbf{T}_w(\mathbf{L})\,\sigma\left(\sum_{j=0}^{W-1}\mathbf{T}_j(\mathbf{L})\mathbf{X}\theta_{j,1}\right)\theta_{w,2}\right) \tag{24}$$

In this expression, we are performing approximate spectral convolution by multiplying the feature matrix $\mathbf{X}$ with multiple Chebyshev polynomials of a scaled Laplacian $\mathbf{L} = \mathbf{I} - \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}$. W is the maximum order of Chebyshev polynomial that we consider and θ are the learnable parameters. We refer the reader to (Defferrard et al., 2016) for a more detailed explanation.

To obtain a linear approximation, we can replace the non-linear operator σ in the hidden layers with the identity matrix:

$$\sigma\left(\sum_{\mathbf{w}=0}^{W-1}\sum_{j=0}^{W-1}\mathbf{T}_w(\mathbf{L})\mathbf{T}_j(\mathbf{L})\mathbf{X}\Theta_{w,j}\right)\,. \tag{25}$$

We can learn the parameters via logistic regression by stacking all of the preprocessed $\mathbf{T}_w\mathbf{T}_j\mathbf{X}$ terms into a new augmented matrix $\hat{\mathbf{X}}$, thereby multiplying the feature dimensionality by a factor of $(W)^\ell$.

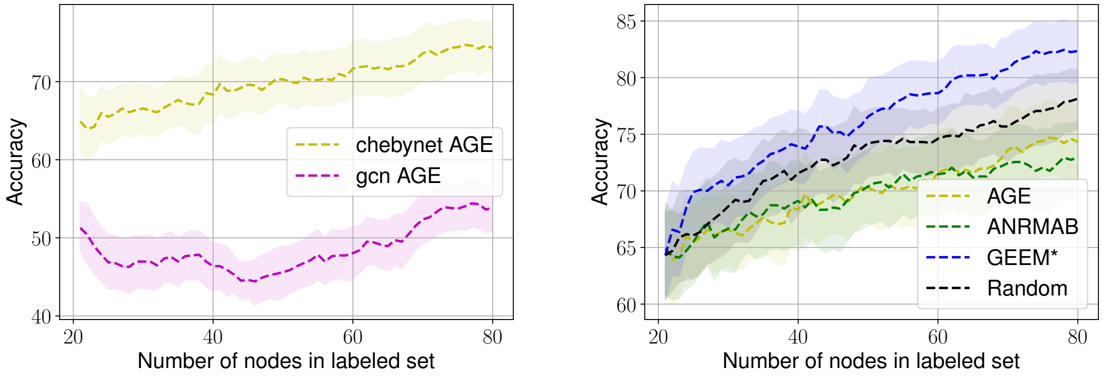
4.2. Experiment

We provide the results of an experiment for a dataset where the first order approximation made by the GCN graph filters is significantly outperformed by the Chebynet. WebKB is a webpages dataset where the node features are representations of webpage content and an edge is added when a website links another. Table 1 provides the statistics of the dataset. We set the number of layers of the GEEM Chebynet to 1 to reduce the time complexity and use $\mathbf{w} = 3$ polynomials because this was the number used in (Defferrard et al., 2016). For the baselines, we also replace the GCN employed in **AGE** and **ANRMAB** by a Chebynet using the same default hyperparameters as described in the main paper. All active learning algorithms commence with the same randomly chosen 21 initial nodes. This value was chosen to give an initial classification accuracy of approximately 65 percent.

Figure 1 depicts the performance of the algorithms. Figure 1(a) highlights how important it is to use a Chebynet-based classifier rather than GCN. In Figure 1(b), we see that AGE and ANRMAB perform similarly and both are outperformed by random selection. This is an indication that the default hyperparameters are not a good match for this new dataset or the Chebynet architecture. In the active learning setting, there is no validation set available to search for better parameters. The GEEM algorithm that incorporates linear Chebynet logistic regression significantly outperforms random selection.

Table 1. Statistics of webKB.

Dataset	#Classes	#Features	#Nodes	#Edges	Edge density
WebKB	5	1,703	251	450	0.03%



(a)

(b)

Figure 2. Experiment 1 for the WebKB dataset. Initial labeled set of 8% nodes a) GCN vs Chebynet AGE. The chebynet version clearly outperforms. b) Comparison of active learning algorithms with Chebynet-based architecture

Table 2. Experiment 1 for WebKB dataset. Average accuracy at different budgets. Asterisks indicate that a Wilcoxon ranking test showed a significant difference (at 5% significance level) between the marked method and the best performing baseline.

budget b	**0**	**1**	**10**	**30**	60
GEEM*	64.3	**66.6**	**71.2**	**77.0***	**82.4***
Random	64.3	64.7	69.2	74.3	78.1
AGE	**64.9**	63.9	66.3	70.0	74.3
ANRMAB	**64.9**	64.2	66.6	70.7	72.9

References

Defferrard, M., Bresson, X., and Vandergheynst, P. Convolutional neural networks on graphs with fast localized spectral filtering. In *Proc. Advances in Neural Information Processing Systems*, pp. 3844–3852, Barcelona, Spain, Dec. 2016.

Jun, K. and Nowak, R. Graph-based active learning: A new look at expected error minimization. In *Proc. IEEE Global Conf. Signal and Information Proc.*, pp. 1325–1329, Dec. 2016.

Kipf, T. and Welling, M. Semi-supervised classification with graph convolutional networks. In *Proc. Int. Conf. Learning Representations*, Toulon, France, Apr. 2017.

Sivan, H., Gabel, M., and Schuster, A. Online linear models for edge computing. In *Proc. Eur. Conf. Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD)*, 2019.

Wu, F., Souza, A., Zhang, T., Fifty, C., Yu, T., and Weinberger, K. Simplifying graph convolutional networks. In *Proc. Int. Conf. Machine Learning*, pp. 6861–6871, Long Beach, CA, US, Jun. 2019.