ICLR

Google Research

Learning from Label Proportions: Bootstrapping Supervised Learners via Belief Propagation





Shreyas Havaldar*,[1], Navodita Sharma*,[1], Karthikeyan Shanmugam[1], Shubhi Sareen[2], Aravindan Raghuveer[1]

{shreyasjh, navoditasharma, karthikeyanvs, shubhisareen, araghuveer}@google.com

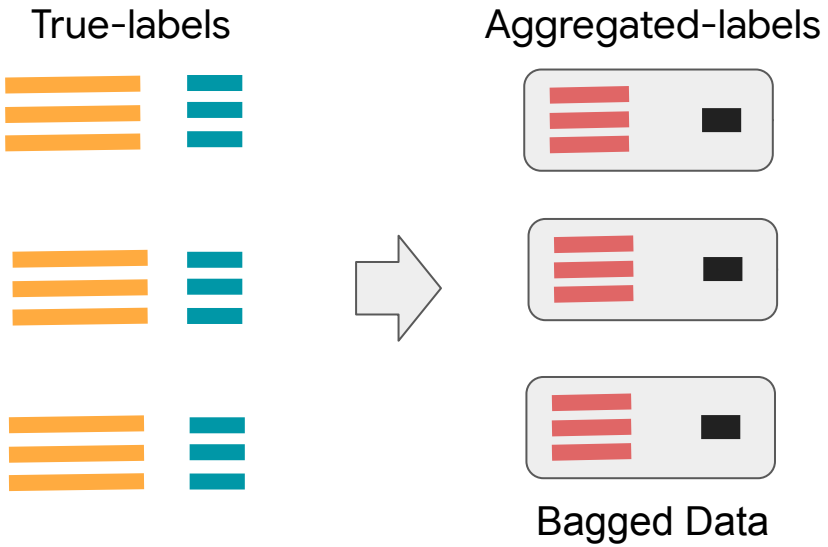
[1]Google Research India, [2]Google India

* denotes equal Contribution

Training a weakly-supervised model, to test on unseen instance data

Learning from Label Proportions

Each bag contains the *average* label of its constituent instances.



Traditional Supervised Learning Data Setup -> Features + Labels

LLP Data Setup -> Groups of features (Bags) + Aggregated Labels

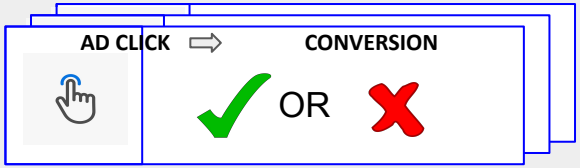
Weak Supervision

Motivation for our work

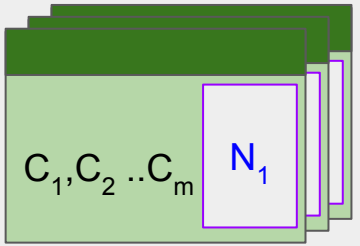
- Train a weakly supervised models based on bag label information but to make predictions on instance data.
- Only have semi supervised data in terms of bag label proportions, to learn from.
- Increasingly important problem of training models privately and within the stricter regulations with respect to data. [1] [2]

[1] Cabral, Luis, et al. "The EU digital markets act: a report from a panel of economic experts." Cabral, L., Haucap, J., Parker, G., Petropoulos, G., Valletti, T., and Van Alstyne, M., The EU Digital Markets Act, Publications Office of the European Union, Luxembourg (2021).
[2] Regulation, General Data Protection. "General data protection regulation (GDPR)." Intersoft Consulting, Accessed in October 24.1 (2018).

But what about real life?



BEFORE

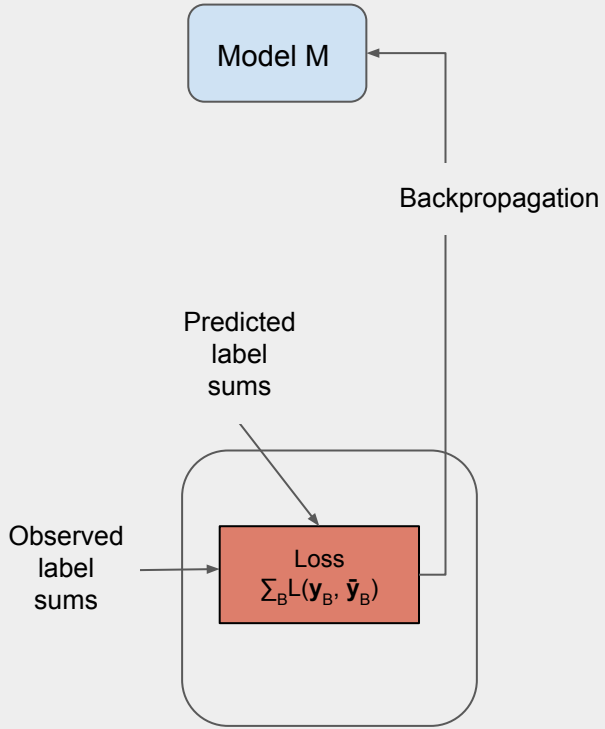


AFTER

Baseline Method

DLLP [1]

- True and predicted label proportions $\mathbf{y}_B$ and $\bar{\mathbf{y}}_B$ for each bag B.
- Loss = $\Sigma_B L(\mathbf{y}_B, \bar{\mathbf{y}}_B)$
- Update model weights.

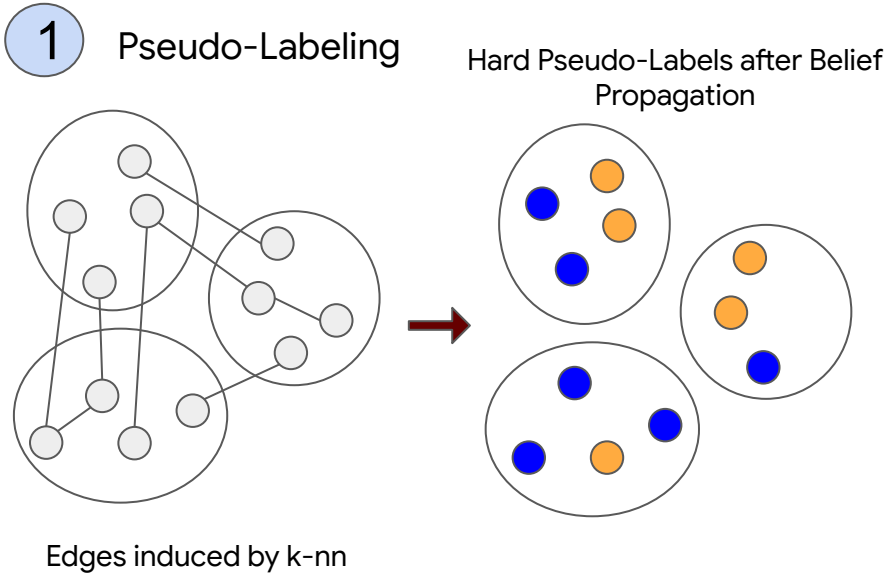


[1] E. M. Ardehaly and A. Culotta. Co-training for demographic classification using deep learning from label proportions. In ICDMW, 2017.

Bootstrapping Supervised Learners via Belief Propagation

Step 2: Use Supervised Learning over features and pseudo-labels.

Step 1: Find Pseudo-labels for instances. We use the aggregated labels and covariate information to do this.

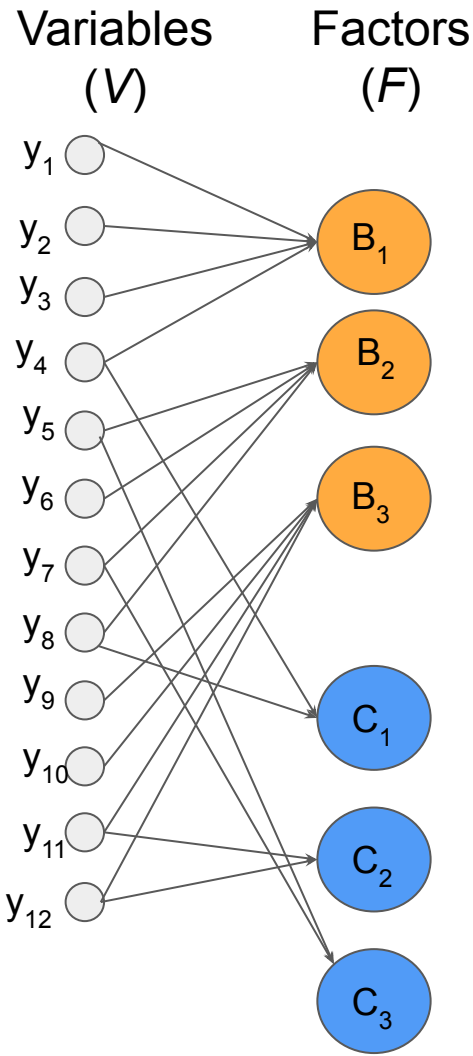


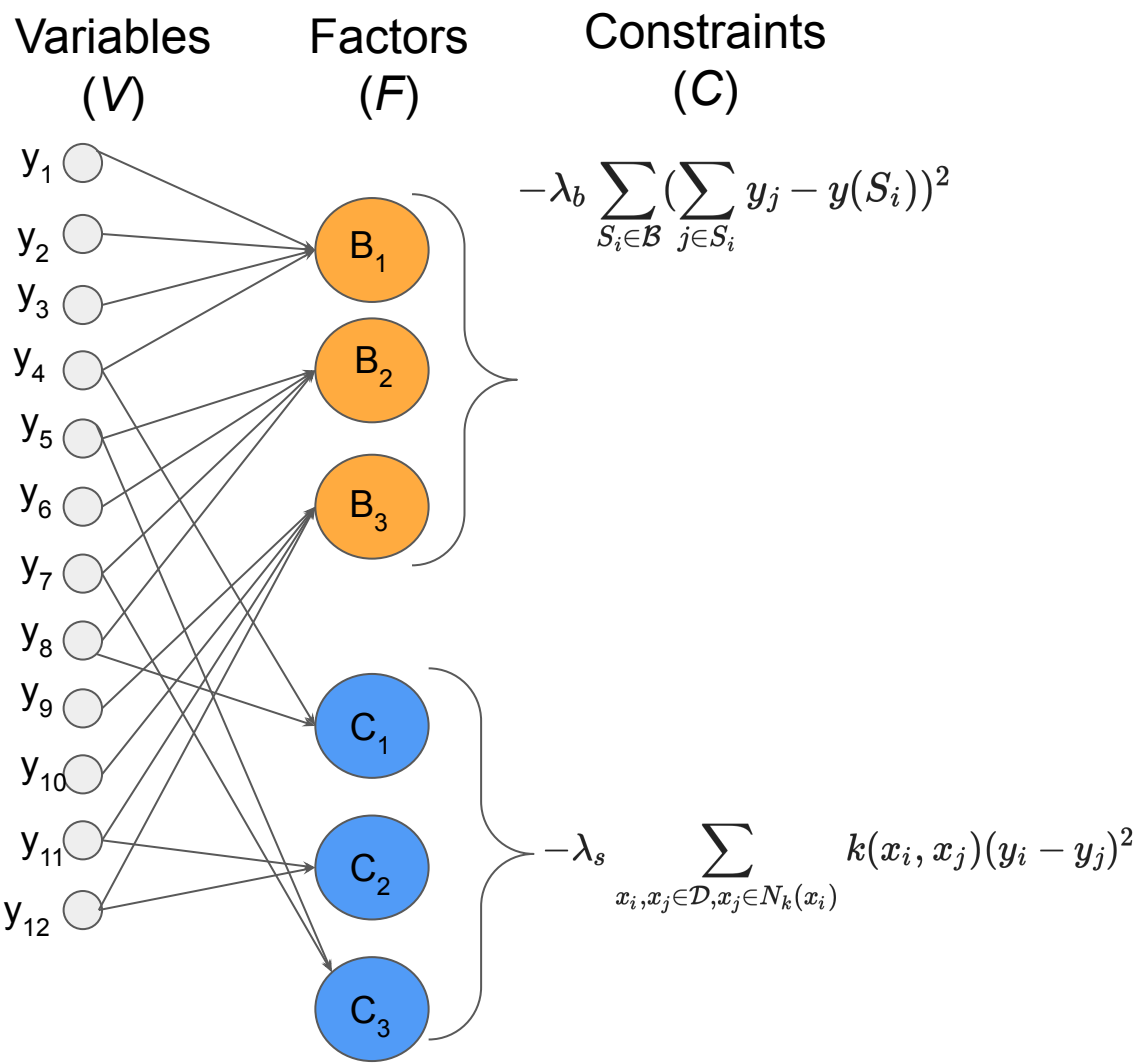
Step 1

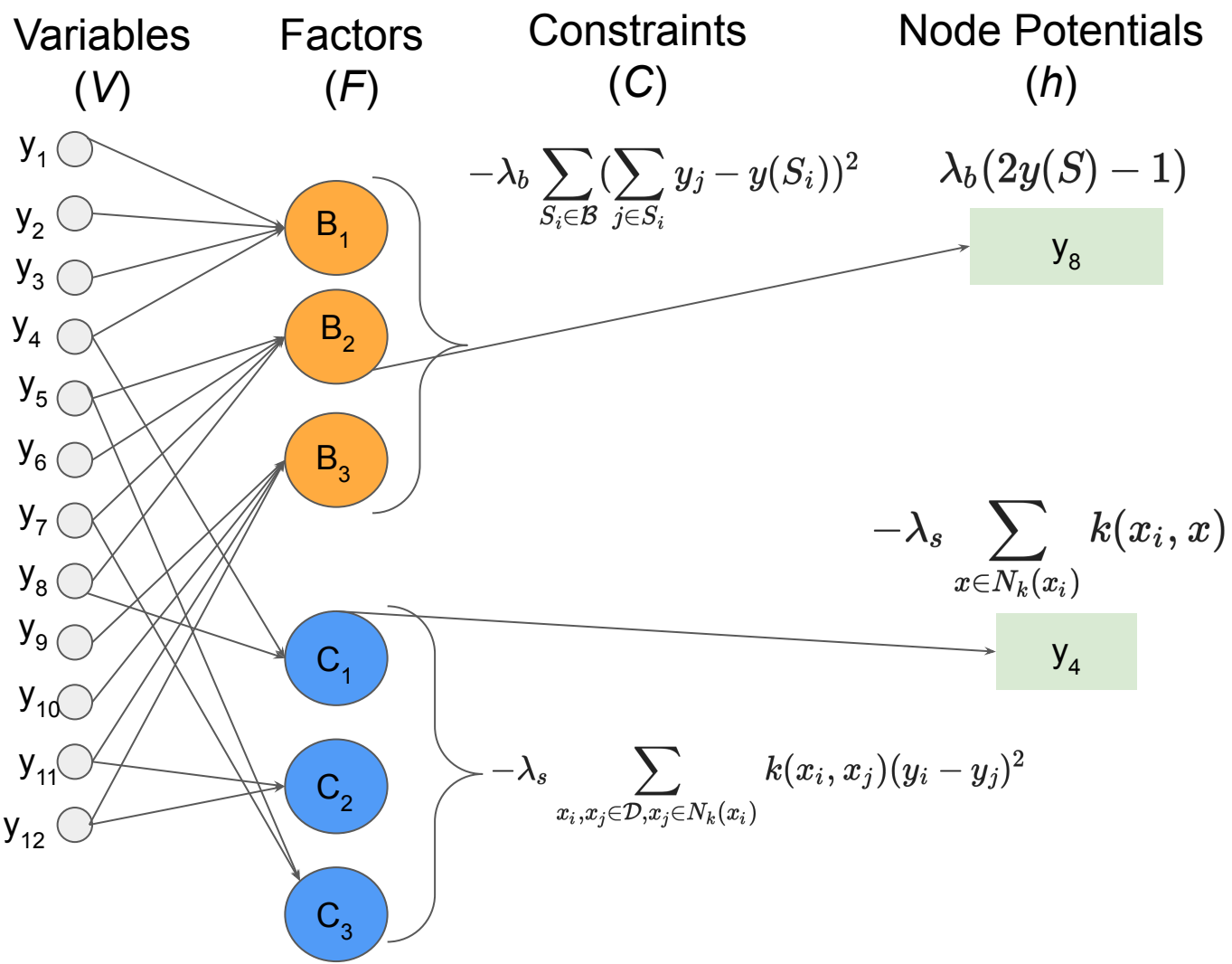
Obtaining *Pseudo-Labels* through Belief Propagation (*BP*)

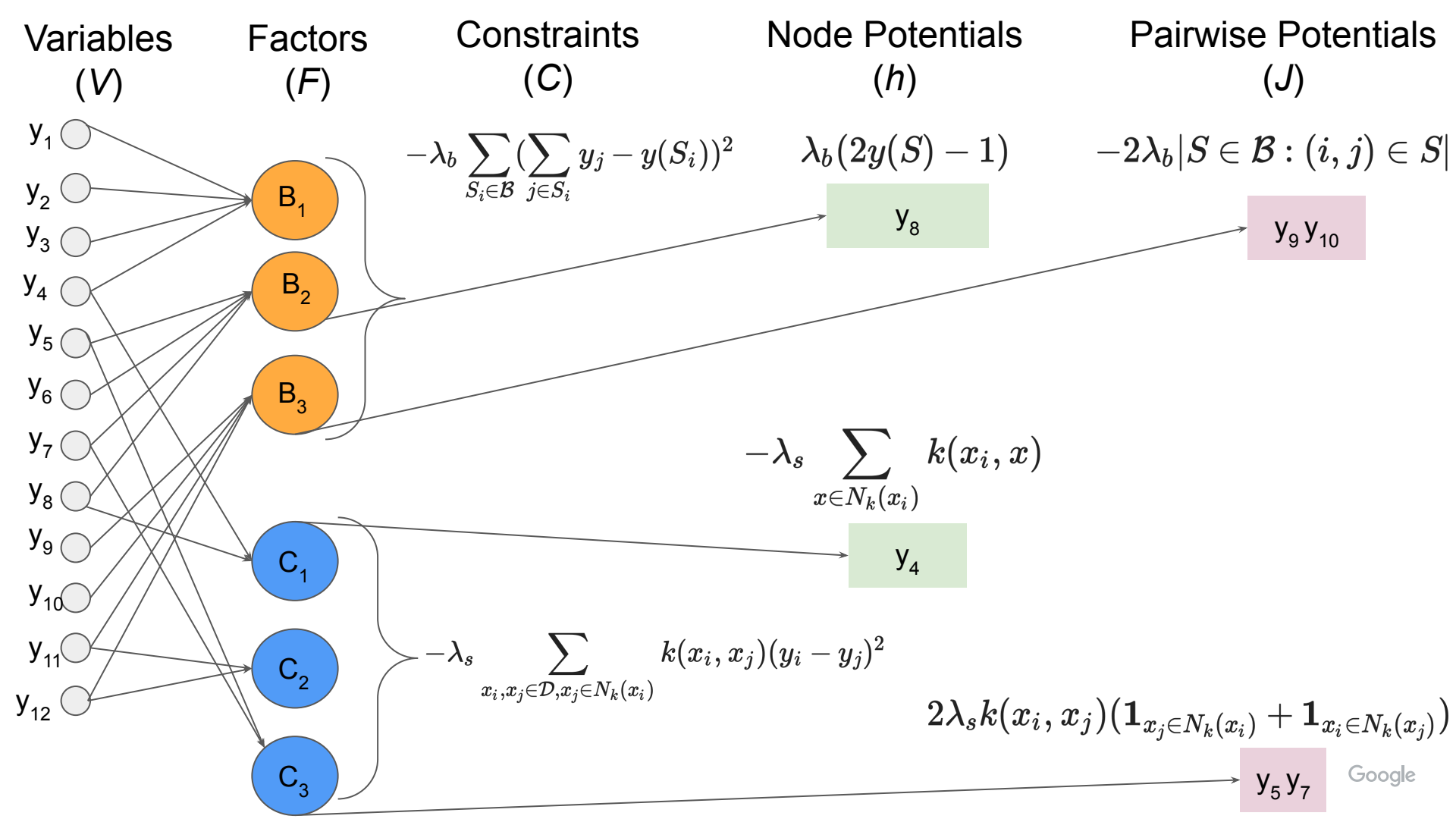
Variables
(V)

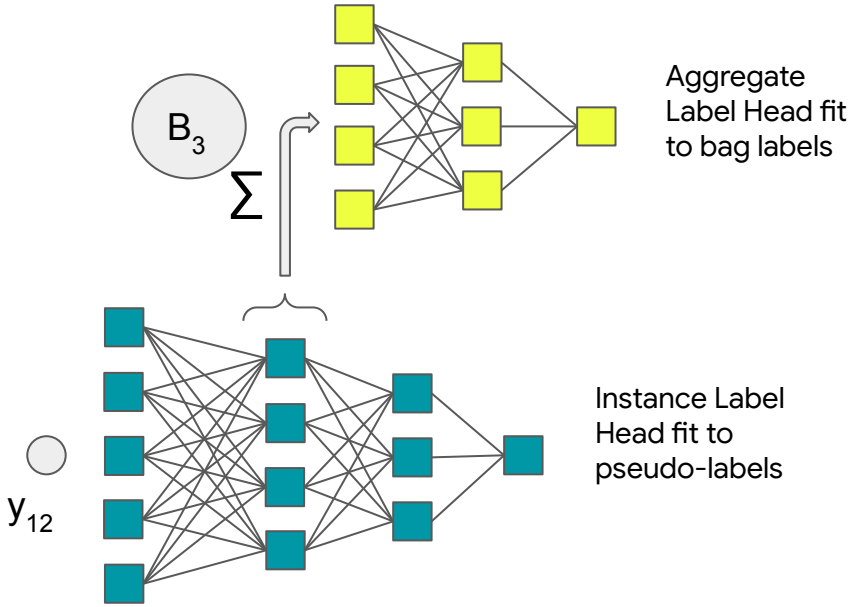
- y_1
- y_2
- y_3
- y_4
- y_5
- y_6
- y_7
- y_8
- y_9
- y_{10}
- y_{11}
- y_{12}


Variables
(V)
Factors
(F)
y1
y2
y3
y4
y5
y6
y7
y8
y9
y10
y11
y12
B1
B2
B3
C1
C2
C3


Variables
(V)
Factors
(F)
Constraints
(C)
y1
y2
y3
y4
y5
y6
y7
y8
y9
y10
y11
y12
B1
B2
B3
C1
C2
C3
$-\lambda_b \sum_{S_i \in \mathcal{B}} (\sum_{j \in S_i} y_j - y(S_i))^2$
$-\lambda_s \sum_{x_i, x_j \in \mathcal{D}, x_j \in N_k(x_i)} k(x_i, x_j)(y_i - y_j)^2$

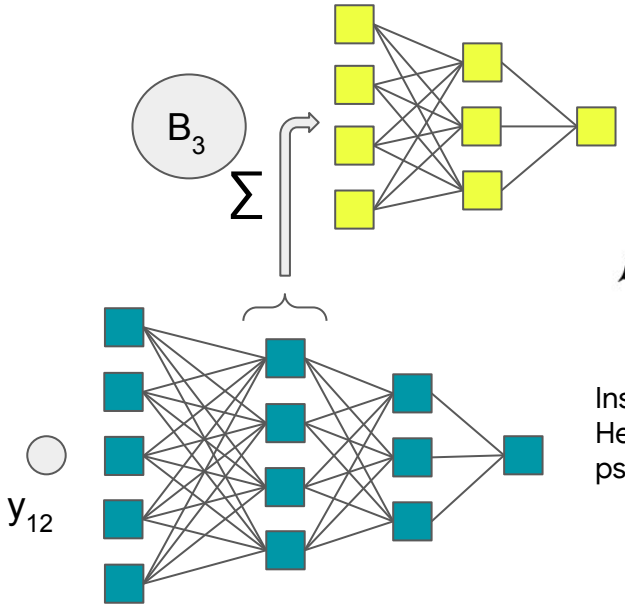

Variables
(V)
Factors
(F)
Constraints
(C)
Node Potentials
(h)
$-\lambda_b \sum_{S_i \in \mathcal{B}} (\sum_{j \in S_i} y_j - y(S_i))^2$
$\lambda_b(2y(S)-1)$
$-\lambda_s \sum_{x \in N_k(x_i)} k(x_i, x)$
$-\lambda_s \sum_{x_i, x_j \in \mathcal{D}, x_j \in N_k(x_i)} k(x_i, x_j)(y_i - y_j)^2$
y1
y2
y3
y4
y5
y6
y7
y8
y9
y10
y11
y12
B1
B2
B3
C1
C2
C3
y8
y4


Variables
(V)
Factors
(F)
Constraints
(C)
Node Potentials
(h)
Pairwise Potentials
(J)
y1
y2
y3
y4
y5
y6
y7
y8
y9
y10
y11
y12
B1
B2
B3
C1
C2
C3
$-\lambda_b \sum_{S_i \in \mathcal{B}} (\sum_{j \in S_i} y_j - y(S_i))^2$
$\lambda_b(2y(S)-1)$
$-2\lambda_b |S \in \mathcal{B} : (i,j) \in S|$
y8
y9 y10
$-\lambda_s \sum_{x \in N_k(x_i)} k(x_i, x)$
y4
$-\lambda_s \sum_{x_i, x_j \in \mathcal{D}, x_j \in N_k(x_i)} k(x_i, x_j)(y_i - y_j)^2$
$2\lambda_s k(x_i, x_j)(\mathbf{1}_{x_j \in N_k(x_i)} + \mathbf{1}_{x_i \in N_k(x_j)})$
y5 y7
Google



Step 2

Embedding Refinement
Leveraging Pseudo Labels

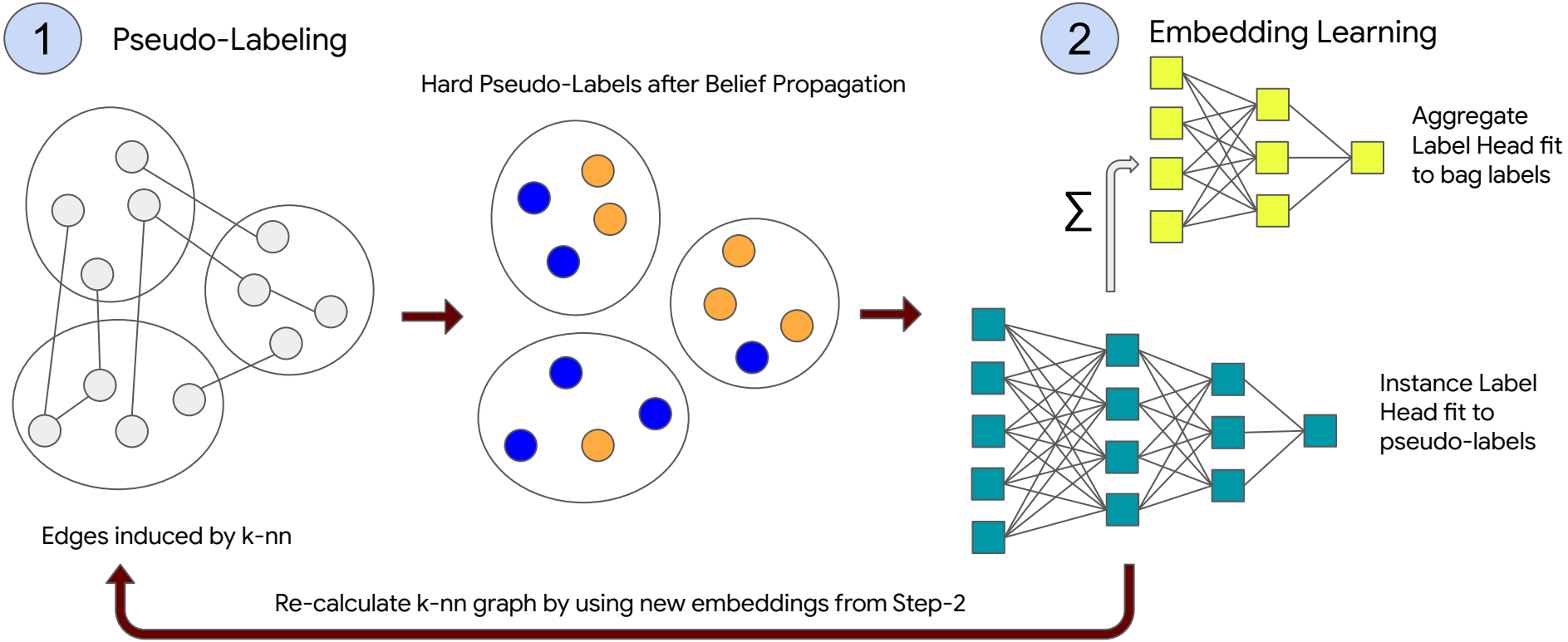


Aggregate Label Head fit to bag labels

$$\mathcal{L}_{\text{Agg-Emb}}(S) = \sum_{i \in S} \text{CE}\left(f_L(x_i), \{y_i^0\}\right) + \lambda_a \text{CE}\left(g_L(S), \frac{y(S)}{|S|}\right)$$

Instance Label Head fit to pseudo-labels

The whole picture



Our Improvements

Improvements

15% Standard UCI classification datasets.

0.8% Large Challenging Criteo Data

7% Image Datasets

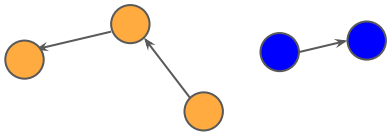
Dataset:	Marketing					
Bag Size:	8	32	128	512	1024	2048
DLLP	84.49 (0.50)	82.65 (0.94)	79.69 (2.03)	70.36 (0.64)	66.39 (2.43)	65.60 (3.21)
EasyLLP	83.63 (0.14)	82.87 (0.72)	75.05 (3.29)	68.97 (2.76)	50.23 (1.21)	50.12 (0.35)
GenBags	85.26 (0.42)	83.15 (0.34)	79.74 (0.50)	69.29 (0.92)	64.82 (3.10)	58.43 (4.31)
Ours-Itr-1	85.76 (0.06)	84.18 (0.03)	**82.71 (0.24)**	77.71 (0.06)	80.56 (0.35)	78.63 (0.63)
Ours-Itr-2	**86.26 (0.01)**	**84.33 (0.05)**	82.46 (0.05)	**81.68 (0.15)**	**81.66 (0.41)**	**81.01 (0.72)**

Dataset:	CIFAR-S					
Bag Size:	8	32	128	512	1024	2048
DLLP	**93.87 (0.11)**	**92.12 (0.24)**	**88.63 (0.51)**	79.58 (1.34)	52.01 (8.56)	57.21 (6.50)
GenBags	92.36 (0.50)	90.10 (0.39)	86.78 (0.33)	82.69 (1.01)	68.45 (3.79)	60.43 (4.49)
EasyLLP	85.54 (1.006)	74.79 (2.17)	65.26 (3.51)	61.57 (9.88)	62.46 (5.21)	52.32 (3.04)
LLP-FC	*	85.58 (0.31)	80.59 (0.56)	75.62 (1.21)	65.75 (2.36)	63.76 (1.26)
LLP-VAT	90.10 (0.49)	83.20 (0.16)	64.76 (3.06)	^	^	^
Ours-Itr-1	93.53 (0.03)	91.17 (0.03)	88.17 (0.19)	82.97 (0.33)	**74.45 (0.58)**	71.01 (1.11)
Ours-Itr-2	93.64 (0.01)	91.31 (0.03)	88.31 (0.01)	**84.30 (0.02)**	&	**71.17 (2.03)**

Dataset:	Criteo		
Bag Size:	8	32	128
DLLP	74.11 (0.09)	72.86 (0.01)	**70.99 (0.01)**
EasyLLP	70.77 (0.92)	68.42 (0.62)	62.87 (1.50)
GenBags	73.34 (0.01)	71.32 (0.77)	70.39 (0.46)
Ours-Itr-1	74.96 (0.01)	73.36 (0.03)	70.45 (0.05)
Ours-Itr-2	**74.97 (0.01)**	**73.39 (0.01)**	70.81 (0.01)

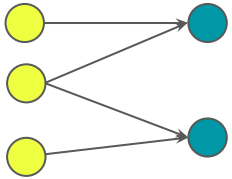
Ablations

1NN Closeness to kNN



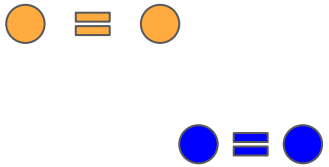
1NN Neighbour Graph achieves most of the gains relative to the kNN for the pseudo-labelling step.

Tree Structure of Factor Graph



Bi-partite graph between instances and factor nodes are tree like for 1NN, favouring BP Convergence.

Covariate Information Essential



We lose upto **18% AUROC** if we drop 90% of the covariate factors.

Efficient Running Time

Feasibility on larger bags and datasets due to only O(m(B+k)) pairwise terms.

Privacy Guarantees

Larger bags -> better privacy
Utilize the better privacy utility degradation tradeoffs

Goodness of Pseudo Labels

Good ordering information and the effect of high quality pseudo labels is reflected in the downstream performance.

Conclusion

- We have provided a highly generalizable algorithm to perform efficient learning from label proportions.
- We utilised Belief Propagation on parity like constraints derived from covariate information and bag level constraints to obtain pseudo-labels.
- We show extensive experimental comparisons against several SOTA baselines across various datasets of different types.
- We perform approximate convergence analysis for our algorithm providing theoretical backing for our strong empirical results.

What’s next?

- Explore alternate energy potentials for the Gibbs distribution.
- Why such a simple proposition like BP works on such a scale efficiently converging to marginals proving highly useful in supervised learning even with 1-NN based covariate information.
- A complete theoretical understanding behind the success of BP for the target task.
- Extending this to learning from diverse input sources.

Thank You!

Questions?

{shreyasjh, navoditasharma}@google.com


That's all Folks!