f_{BGD}: Learning Embeddings From Positive Unlabeled Data with BGD

A Dot Product Structures in Tensor Models

(1) For the Pairwise Interaction Tensor Factorization (Rendle and Schmidt-Thieme, 2010) (Note that here $\mathbf{V}^X$, $\mathbf{V}^Y$ and $\mathbf{V}^H$ are shared for pairwise interactions, while in the original paper they are independent.):

$$\hat{r}_{xy} = \sum_{d=1}^{f} v^X_{j,d} v^H_{j',d} v + \sum_{d=1}^{f} v^X_{j,d} v^Y_{j'',d} + \sum_{d=1}^{f} v^H_{j',d} v^Y_{j'',d} \tag{1}$$

we have $g = f + 1$ and

$$\begin{aligned} p_{xh,d} &= v^X_{j,d} + v^H_{j',d}\ , \quad q_{y,d} = v^Y_{j'',d} \\ p_{xh,f+1} &= \sum_{d=1}^{f} v^X_{j,d} v^H_{j',d}\ , \quad q_{y,f+1} = 1 \end{aligned} \tag{2}$$

(2) For a tensor factorization (TF) — rank-f Canonical Polyadic Decomposition (Bailey and Aeron, 2017):

$$\hat{r}_{xy} = \sum_{d=1}^{f} v^X_{j,d} v^H_{j',d} v^Y_{j'',d} \tag{3}$$

we have $g = f$ and

$$p_{xh,d} = v^X_{j,d} v^H_{j',d}\ , \quad q_{y,d} = v^Y_{j'',d} \tag{4}$$

where the gradient, e.g., $\nabla_{v^X_{j,d}} p_{xh,d}$ is $v^H_{j',d}$. Note we assume that $v^X_{j,d}$ and $v^H_{j',d}$ belong to x-related parameters, while $v^Y_{j'',d}$ is a parameter related to y.

B Learning of f_{BGD}

Algorithm 1 summarizes the accelerated algorithm for f_{BGD}. We define the $S^p_{dd'}$ and S^p_d caches as $S^p_{dd'} = \sum_{x \in X} p_{x,d} p_{x,d'}$ and $S^p_d = \sum_{x \in X} p_{x,d}$ respectively. The gradients w.r.t. $\theta^Y \in \Theta^Y$ is given by

$$\nabla_{\theta^Y} \widetilde{J}_A(\theta) = 2 \sum_{d=1}^{g} \sum_{d'=1}^{g} S^p_{dd'} \sum_{y \in Y} \alpha^-_y q_{y,d'} \nabla_\theta q_{y,d} - 2r^- \sum_{d=1}^{g} S^p_d \sum_{y \in Y} \alpha^-_y \nabla_\theta q_{y,d} \tag{5}$$

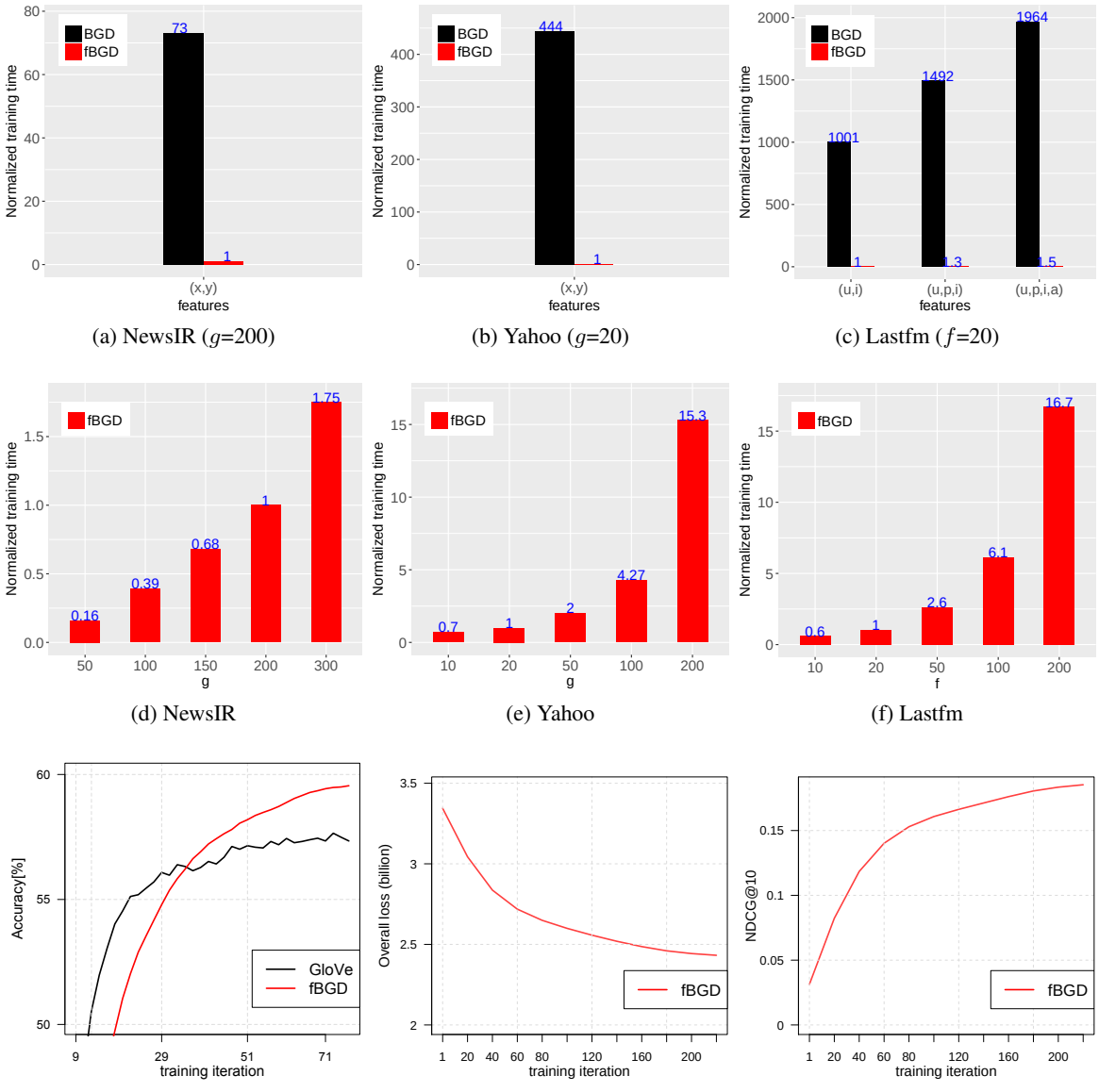
Note that Line 5-10 and 21 can be omitted when optimizing the basic dot product function; g' in Line 18 equals to g and $g - 2$ (i.e., f) for the basic dot product function and SVDFeature (Eq. (11)) respectively.

C Runtime Results

References

E. Bailey and S. Aeron. Word embeddings via tensor factorization. *arXiv preprint arXiv:1704.02686*, 2017.

S. Rendle and L. Schmidt-Thieme. Pairwise interaction tensor factorization for personalized tag recommendation. In *WSDM*, 2010.



(a) NewsIR (g=200) (b) Yahoo (g=20) (c) Lastfm (f=20)

(d) NewsIR (e) Yahoo (f) Lastfm

(g) NewsIR (h) Yahoo (i) Lastfm

Figure 1: (a), (b) and (c) show the runtime per iteration (f_{BGD} vs BGD). One unit in (a)(d), (b)(e) and (c)(f) is 388, 165, 26 seconds respectively. (d), (e) and (f) show the runtime change (per iteration) by increasing embedding size. (g), (h) and (i) show the convergence behavior reflected by the training loss or predicted accuracy.

Algorithm 1 Generic f_{BGD} Learning Algorithm

1: **Input:** P, X, Y, Cache vectors $\mathbf{s}^q$, $\mathbf{s}^p$, Cache matrices $\mathbf{E}$, $\mathbf{Q}$, $\mathbf{S}^q$, $\mathbf{S}^p$;
2: **Output:** Θ
3: Initialize $\Theta \sim \mathcal{N}(0, 0.01)$:
4: **for** $e = 1, ...,$ maxiter **do**
5: **for** $d \in \{1, .., g\}$ **do**
6: **for** $x \in X$ **do**
7: Compute p_{xd}, store in $\mathbf{E}$ ($\mathbf{E} \in \mathbb{R}^{|X| \times g}$)
8: **end for**
9: Repeat line 6 to 8 for $y \in Y$
10: **end for**
11: **for** $d \in \{1, .., g\}$ **do**
12: Compute S^q_d, store in $\mathbf{s}^q$ ($\mathbf{s}^q \in \mathbb{R}^g$)
13: **for** $d' \in \{1, .., g\}$ **do**
14: Compute $S^q_{dd'}$, store in $\mathbf{S}^q$ ($\mathbf{S}^q \in \mathbb{R}^{g \times g}$)
15: **end for**
16: **end for**
17: **for** $j \in \{1, .., pX\}$ **do**
18: **for** $d \in \{1, .., g'\}$ **do**
19: Compute $\nabla_{\theta^X} J_A(\Theta)$,$\nabla_{\theta^X} J_p(\Theta)$
20: Update θ^X as in Eq.(21)
21: Update $p_{x,d}$ as in Eq.(20)
22: **end for**
23: **end for**
24: Repeat line 11 to 23 for updating θ^Y
25: **end for**