Frequency-aware SGD for Efficient Embedding Learning with Provable Benefits

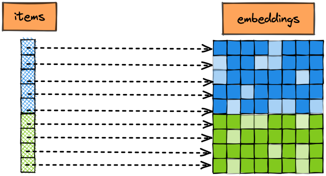
Yan Li[1], Dhruv Choudhary[2], Xiaohan Wei[2], Baichuan Yuan[2], Bhargav Bhushanam[2], Tuo Zhao[1], Guanghui (George) Lan[1]

[1]Georgia Institute of Technology, [2]Meta

ICLR 2022

Problem Formulation

Embedding Learning Problems



$$\min_{\Theta \in \mathbb{R}^{N \times d}} f(\Theta) = \mathbb{E}_{(i,j)\sim\mathcal{D}} \left[\ell(\theta_i, \theta_j; y_{ij})\right] = \sum_{i \in U, j \in V} D(i,j)\ell(\theta_i, \theta_j; y_{ij})$$

- $D(i,j)$: occurrence prob. of (i,j) pair
- y_{ij}: interaction label
- θ_i, θ_j: embedding vector of item i, j, respectively
- N: $\#$ items
- d: embedding dimension

How to learn embedding efficiently?

Practices & Intuitions

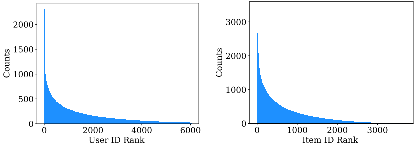
Standard Practices

$$\min_{\Theta \in \mathbb{R}^{N \times d}} f(\Theta) = \mathbb{E}_{(i,j)\sim\mathcal{D}}\left[\ell(\theta_i, \theta_j; y_{ij})\right] = \sum_{i \in U, j \in V} D(i,j)\ell(\theta_i, \theta_j; y_{ij})$$

- Popular choices of opt. methods: Adagrad/Adam
- SGD gives significantly (incomparably) worse performance
  - Liu et al., Understanding the difficulty of training transformers, 2020
  - Zhang et al., Why are adaptive methods good for attention models? 2019

What causes this gap? Any intuition?

- Adaptive methods use larger learning rates for infrequent items



(a) # Users, Movielens **(b)** # Item, Movielens

Try to learn infrequent items faster ..

Theory - Practice Gap

Theory seems hard to catch up

Convergence rate of Adaptive methods (Adagrad/Adam) compared to SGD:

- Convex setting:
  - Duchi et al. '11: better dimensional dependency
- Nonconvex setting:
  - Ward et al. '18, Defossez et al. '20, Chen et al. '18, Zhou et al. '18:
    - ⋆ hardly matches SGD
    - ⋆ improvement relies on strong assumptions

Emedding learning is often nonconvex, can we reconcile theory-practice gap?

Frequence-aware SGD

SGD - but adaptive to item frequency

Algorithm Frequency-aware SGD

Input: Total iteration number T, token frequency $\{p_k\}_{k\in X}$, and learning rate schedule $\{\eta_k^t\}_{k\in X, t\in[T]}$ specified by $\eta_k^t = \min\left\{1/(4L), \alpha/\sqrt{Tp_k}\right\}$.

Initialize: $\Theta^0 \in \mathbb{R}^{N\times d}$, sample $\tau \sim \mathrm{Unif}([T])$,

for $t = 0, \ldots \tau$ **do**

(1) Sample $(i_t, j_t) \sim \mathcal{D}$, calculate $g_{i_t}^t = \nabla_{\theta_{i_t}} \ell(\theta_{i_t}, \theta_{j_t}; y_{i_t,j_t})$, $g_{j_t}^t = \nabla_{\theta_{j_t}} \ell(\theta_{i_t}, \theta_{j_t}; y_{i_t,j_t})$

(2) Update parameters

$$\theta_{i_t}^{t+1} = \theta_{i_t}^t - \eta_{i_t}^t g_{i_t}^t, \quad \theta_i^{t+1} = \theta_i^t, \quad \forall i \in U, i \neq i_t$$

$$\theta_{j_t}^{t+1} = \theta_{j_t}^t - \eta_{j_t}^t g_{j_t}^t, \quad \theta_j^{t+1} = \theta_j^t, \quad \forall j \in V, j \neq j_t$$

end for

Output: Θ^τ

⋆ Use larger learning rates for infrequent items – but with an explicit rule!

Convergence of FA-SGD v.s. SGD

Theorem (FA-SGD)

Take $\alpha = \sqrt{(f(\Theta^0) - f^*) \,/\, \left(L \sum_{l \in X} p_l \sigma_l^2\right)}$ *in FA-SGD, we have*

$$\mathbb{E}\|\nabla f_k^\tau\|^2 = \mathcal{O}\left(\frac{L\left(f(\Theta^0)-f^*\right)}{T} + \frac{\sqrt{p_k}\sqrt{\sum_{l\in X} p_l \sigma_l^2 (f(\Theta^0)-f^*)L}}{\sqrt{T}}\right), \quad \forall k \in X$$

Theorem (Standard SGD)

Take learning rate policy to be $\eta_k^t = \min\left\{\frac{1}{4L}, \frac{\alpha}{\sqrt{T}}\right\}$, *where* T *denotes the total number of iterations, and* $\alpha = \sqrt{\frac{f(\Theta^0)-f^*}{L\sum_{l\in X} p_l^2 \sigma_l^2}}$, *we have*

$$\mathbb{E}\|\nabla f_k^\tau\|^2 = \mathcal{O}\left(\frac{L\left(f(\Theta^0)-f^*\right)}{T} + \frac{\sqrt{\sum_{l\in X} p_l^2 \sigma_l^2 (f(\Theta^0)-f^*)L}}{\sqrt{T}}\right), \quad \forall k \in X$$

Implications?

Convergence of FA-SGD v.s. SGD

Theorem (FA-SGD)

Take $\alpha = \sqrt{(f(\Theta^0) - f^*) / (L \sum_{l \in X} p_l \sigma_l^2)}$ *in FA-SGD, we have*

$$\mathbb{E}\|\nabla f_k^\tau\|^2 = \mathcal{O}\left(\frac{L\big(f(\Theta^0)-f^*\big)}{T} + \frac{\sqrt{p_k}\sqrt{\sum_{l\in X} p_l \sigma_l^2 (f(\Theta^0)-f^*)L}}{\sqrt{T}}\right), \quad \forall k \in X$$

Theorem (Standard SGD)

Take learning rate policy to be $\eta_k^t = \min\left\{\frac{1}{4L}, \frac{\alpha}{\sqrt{T}}\right\}$, *where* T *denotes the total number of iterations, and* $\alpha = \sqrt{\frac{f(\Theta^0)-f^*}{L\sum_{l\in X} p_l^2 \sigma_l^2}}$, *we have*

$$\mathbb{E}\|\nabla f_k^\tau\|^2 = \mathcal{O}\left(\frac{L\big(f(\Theta^0)-f^*\big)}{T} + \frac{\sqrt{\sum_{l\in X} p_l^2 \sigma_l^2 (f(\Theta^0)-f^*)L}}{\sqrt{T}}\right), \quad \forall k \in X$$

Implications?

Convergence of FA-SGD v.s. SGD

Theorem (FA-SGD)

Take $\alpha = \sqrt{(f(\Theta^0) - f^*) \,/\, (L \sum_{l \in X} p_l \sigma_l^2)}$ *in FA-SGD, we have*

$$\mathbb{E}\|\nabla f_k^\tau\|^2 = \mathcal{O}\left(\frac{L\big(f(\Theta^0) - f^*\big)}{T} + \frac{\sqrt{p_k}\sqrt{\sum_{l \in X} p_l \sigma_l^2 (f(\Theta^0) - f^*) L}}{\sqrt{T}} \right), \quad \forall k \in X$$

Theorem (Standard SGD)

Take learning rate policy to be $\eta_k^t = \min\left\{\frac{1}{4L}, \frac{\alpha}{\sqrt{T}}\right\}$, *where* T *denotes the total number of iterations, and* $\alpha = \sqrt{\frac{f(\Theta^0) - f^*}{L \sum_{l \in X} p_l^2 \sigma_l^2}}$, *we have*

$$\mathbb{E}\|\nabla f_k^\tau\|^2 = \mathcal{O}\left(\frac{L\big(f(\Theta^0) - f^*\big)}{T} + \frac{\sqrt{\sum_{l \in X} p_l^2 \sigma_l^2 (f(\Theta^0) - f^*) L}}{\sqrt{T}} \right), \quad \forall k \in X$$

Implications?

Provable Benefits of FA-SGD

Corollary (Exponential Tail)

Let $U = \{i_n\}_{n=1}^{|U|}$, $V = \{j_m\}_{m=1}^{|V|}$, where i_n denote the user with n-th largest frequency, and j_m denote the item with the m-th largest frequency. Suppose $p_{i_n} \propto \exp(-\tau n), p_{j_m} \propto \exp(-\tau m)$, for some $\tau > 0$. Define U_T as the set of users whose frequencies are within e-factor from the highest frequency. Then given $|U|, |V| \geq \frac{1}{\tau}$, FA-SGD, compared to standard SGD:

(1) *Obtains the same rate of convergence, for the top users U_T and top items V_T;*

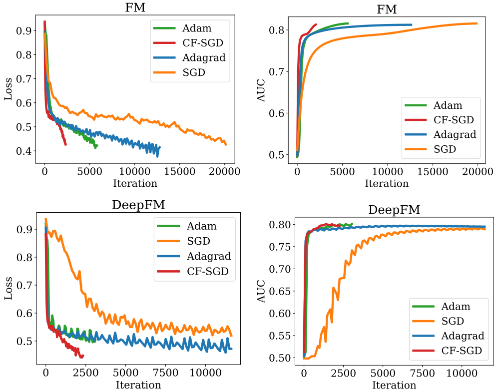
(2) $\mathbb{E}\|\nabla f_{i_n}^{\tau}\|^2$ *can converge faster by a factor of* $\Omega\left\{\exp\left(\tau(n - |U_T|)\right)\right\}$ *for* $i_n \in U \setminus U_T$ *;*

(3) $\mathbb{E}\|\nabla f_{j_m}^{\tau}\|^2$ *can converge faster by a factor of* $\Omega\left\{\exp\left(\tau(m - |V_T|)\right)\right\}$ *for* $j_m \in V \setminus V_T$.

First theoretical speed-up of adaptive methods w.o. algorithmic assumptions

Benchmark Recommendation Task

Movielens-1M

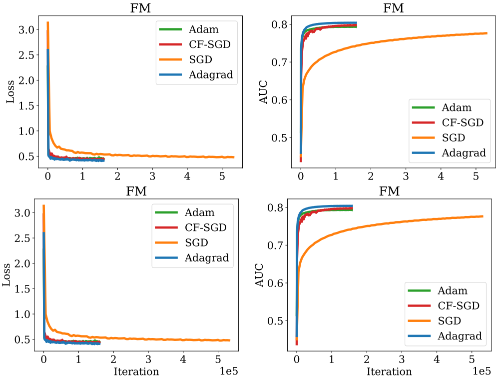


(a) Training loss (b) Validation AUC

Figure: Movielens-1M dataset with FM and DeepFM model. CF-SGD significantly outperforms standard SGD, and is highly competitive against Adam, Adagrad.

Benchmark Recommendation Task

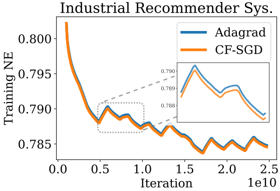
Criteo 1TB Click Logs



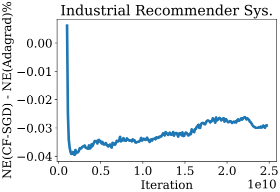
(a) Training loss (b) Validation AUC

Benchmark Recommendation Task

Industrial Recommendation System



(a) Train NE Curve

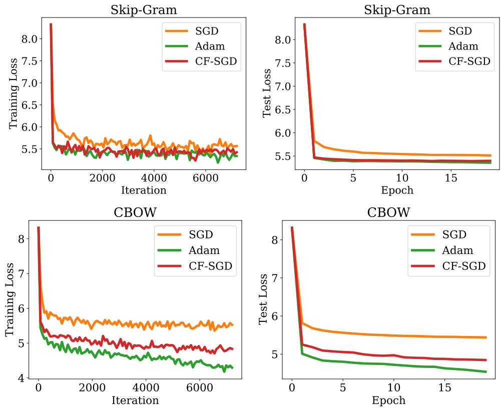


(b) Train NE Diff %

- ~2.5 billion examples per day (*25 billion examples in total*)
- ~ 800 features, with ~100 million average number of tokens per feature
- huge memory savings compared to standard Adagrad/Adam

Learning Word2Vec Embeddings

CBOW & Skip-Gram Models



(a) Training loss **(b)** Testing loss

⋆ Broad applicability of FA-SGD!

Conclusion

- Provable benefits of FA-SGD whenever item/token distribution is imbalanced
- Strong empirical performance
- Memory efficient

Please check out our paper!