FREQUENCY-AWARE SGD FOR EFFICIENT EMBEDDING LEARNING WITH PROVABLE BENEFITS

Anonymous authors
Paper under double-blind review

ABSTRACT

Embedding learning has found widespread applications in recommendation systems and natural language modeling, among other domains. To learn quality embeddings efficiently, adaptive learning rate algorithms have demonstrated superior empirical performance over SGD, largely accredited to their token-dependent learning rate. However, the underlying mechanism for the efficiency of token-dependent learning rate remains underexplored. We show that incorporating frequency information of tokens in the embedding learning problems leads to provably efficient algorithms, and demonstrate that common adaptive algorithms implicitly exploit the frequency information to a large extent. Specifically, we propose (Counter-based) Frequency-aware Stochastic Gradient Descent, which applies a frequency-dependent learning rate for each token, and exhibits provable speed-up compared to SGD when the token distribution is imbalanced. Empirically, we show the proposed algorithms are able to improve or match adaptive algorithms on benchmark recommendation tasks and a large-scale industrial recommendation system, closing the performance gap between SGD and adaptive algorithms, while using significantly lower memory. Our results are the first to show token-dependent learning rate provably improves convergence for non-convex embedding learning problems.

1 INTRODUCTION

Embedding learning describes a problem of learning dense real-valued vector representation for categorical data, often referred to as token (Pennington et al., 2014; Mikolov et al., 2013a;b). Good quality embeddings can capture rich semantic information of tokens, and thus serve as the cornerstone for downstream applications (Santos et al., 2020). Due to their significant impact on model performance and large memory footprint (21.8% of total parameters for BERT (Devlin et al., 2018), 95% for industrial recommenders in Section 4), how to learn quality embedding vectors efficiently forms an important problem in applications, including recommendation systems and natural language processing.

Empirically, adaptive algorithms (Duchi et al., 2011; Kingma & Ba, 2014; Reddi et al., 2019) have witnessed significant successes, yielding state of the art performance in both industrial-scale recommendation systems and natural language model (Guo et al., 2017; Zhou et al., 2018b; Devlin et al., 2018; Liu et al., 2019). Stochastic gradient descent (SGD), on the other hand, has struggled to keep up, often yielding much slower convergence and low quality models (Liu et al., 2020; Zhang et al., 2019) (see also Figure 2). The sharp contrast on the efficiency of adaptive algorithms and SGD is particularly distinctive, as SGD is the typical choice of optimization algorithms in the other domains of machine learning, such as vision/image related tasks (He et al., 2016; Goyal et al., 2017).

The common belief behind the empirical edge of adaptive learning rate algorithms over SGD is that the former ones exploit sparsity of high dimensional feature. Specifically, a feature in a typical embedding learning problem comes in the form of one/multi-hot encoding of tokens (e.g. workpiece in NLP and user/item in recommendation systems), which leads to a sparse stochastic gradient that has only non-zero values for tokens within the mini-batch. In addition, token distributions of real world data are often highly imbalanced and satisfy the power-law property (Piantadosi, 2014; Celma, 2010; Clauset et al., 2009), and infrequent tokens are widely believed to be more informative to model learning. Thus adaptive algorithms can pick up information from the infrequent tokens more efficiently, as they can schedule a higher learning rate for the infrequent tokens (Duchi et al., 2011).

Despite the appealing intuition, there is a significant theory-practice gap on the empirical superiority of adaptive learning rate algorithms over SGD, and no developed theories can explicitly justify the

previous intuition. Better dimensional dependence of adaptive algorithms has only been shown in the convex setting (Duchi et al., 2011), which hardly generalizes to even the simplest practical models in embedding learning problems (e.g., Factorization Machine, Rendle (2010)), whose loss landscape is non-convex. For non-convex settings, most theoretical efforts have been devoted to analyzing adaptive learning rate algorithms for general non-convex objectives, which yield subpar convergence rate compared to standard SGD (Ward et al., 2018; Défossez et al., 2020; Chen et al., 2018; Zhou et al., 2018a). In fact, the standard SGD has been recently shown to be minimax optimal for non-convex problems (Drori & Shamir, 2020; Arjevani et al., 2019), and thus not improvable in general. Moreover, since adaptive algorithms are only implicitly exploiting frequency information, and if the intuition indeed holds true, one might naturally wonder whether we can instead develop an adaptive learning rate schedule that explicitly depends on frequency information. Motivated by our previous discussions, we raise and aim to address the following questions

Questions

Can we design a frequency-dependent adaptive learning rate schedule? Can we show provable benefits over SGD?

Our contributions. We answer the previous question by showing that token frequency information can be leveraged to design provably efficient algorithms for embedding learning. Specifically,

- We propose **F**requency-**a**ware **S**tochastic **G**radient **D**escent (FA-SGD), a simple modification to standard SGD, which applies a token-dependent learning rate that inversely proportional to the frequency of the token. We also propose a variant, named **C**ounter-based **F**requency-aware **S**tochastic **G**radient **D**escent (CF-SGD), which is able to estimate frequency in an online fashion, much similar to Adagrad (Duchi et al., 2011) and Adam (Kingma & Ba, 2014).
- Theoretically, we show that both FA-SGD and CF-SGD outperform standard SGD for embedding learning problems. Specifically, they are able to significantly improve convergence for learning infrequent tokens, while maintaining convergence speed for frequent tokens. To the best of our knowledge, our proposed algorithms are the first to show provable speed-up over standard SGD for non-convex embedding learning problems. This is in sharp contrast with other popular adaptive learning rate algorithms, whose empirical performance can not be explained by existing theories.
- Empirically, we conduct extensive experiments on benchmark datasets and a large-scale industrial recommendation system. We show that FA/CF-SGD is able to significantly improve over SGD, and improves/matches popular adaptive learning rate algorithms. We also observe the second-order moment maintained by Adagrad and Adam highly correlates with the frequency information, demonstrating intimate connections between adaptive algorithms and the proposed FA/CF-SGD.

1.1 Related Literature

Adaptive algorithms for non-convex problems. There has been a fruitful line of research on analyzing the convergence of adaptive learning rate algorithms in non-convex setting. These results aim to match the convergence rate of standard SGD given by $\mathcal{O}(1/\sqrt{T})$ (Ghadimi & Lan, 2013), however often with additional factor of $\log T$ (Ward et al., 2018; Défossez et al., 2020; Chen et al., 2018; Reddi et al., 2018), or with worse dimension dependence (Zhou et al., 2018a) for smooth problem (assumed by almost all prior works). Moreover, all existing works aim to analyze the convergence for general non-convex problems, ignoring unique data features in embedding learning problems, where adaptive algorithms are most successful. We explicitly take account into the sparsity of stochastic gradient, and token distribution imbalancedness into the design and analysis of our proposed algorithms, which are the keys to better convergence properties.

Adaptive algorithms and SGD. To the best of our knowledge, the study on understanding why adaptive learning rate algorithms outperform SGD is very limited. Zhang et al. (2019) argue that BERT pretraining (Devlin et al., 2018) has heavy-tailed noise, implying unbounded variance and possible non-convergence of SGD. Normalized gradient clipping method is proposed therein and converges for a family of heavy-tailed noise distributions. Our results focus on a different direction by showing that imbalanced token distribution is an important factor that can be leveraged to design more efficient algorithms for embedding learning problems. Our result also does not rely on the noise to be heavy-tailed for the convergence benefits of the proposed FA/CF-SGD to take effect.

Algorithm 1 Frequency-aware Stochastic Gradient Descent

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 (7).
Initialize: $\Theta^0 \in \mathbb{R}^{N\times d}$, sample $\tau \sim \mathrm{Unif}([T])$,
for $t = 0, \dots \tau$ **do**
(1) Sample $(i_t, j_t) \sim \mathcal{D}$, calculate the stochastic gradient

$$g_{i_t}^t = \nabla_{\theta_{i_t}} \ell(\theta_{i_t}, \theta_{j_t}; y_{i_t,j_t}), \;\; g_{j_t}^t = \nabla_{\theta_{j_t}} \ell(\theta_{i_t}, \theta_{j_t}; y_{i_t,j_t})$$

(2) Update parameters

$$\theta_{i_t}^{t+1} = \theta_{i_t}^t - \eta_{i_t}^t g_{i_t}^t, \;\; \theta_i^{t+1} = \theta_i^t, \;\; \forall i \in U, i \neq i_t$$
$$\theta_{j_t}^{t+1} = \theta_{j_t}^t - \eta_{j_t}^t g_{j_t}^t, \;\; \theta_j^{t+1} = \theta_j^t, \;\; \forall j \in V, j \neq j_t$$

end for
Output: Θ^τ

Notations: For a vector/matrix, we use $\|\cdot\|$ to denotes its ℓ_2-norm/Frobenius norm. We use $\|\cdot\|_2$ to denote the spectral norm of a matrix.

2 Problem Setup

We consider an embedding learning problem which aims to learn user and item embeddings through their interactions. We denote U as the set of users, and V as the set of items, and let $X = U \cup V$ denote the union, referred to as tokens throughout the rest of the paper. We assume $|X| = N$, i.e., the total number of user and item is N. For the ease of presentation, we always use letter i to index user set U, letter j to index item set V, and letter k to index the union set X. The embedding learning problem can be abstracted into the following stochastic optimization problem:

$$\min_{\Theta\in\mathbb{R}^{N\times d}} f(\Theta) = \mathbb{E}_{(i,j)\sim\mathcal{D}}\left[\ell(\theta_i, \theta_j; y_{ij})\right] = \sum_{i\in U, j\in V} D(i,j)\ell(\theta_i, \theta_j; y_{ij}). \tag{1}$$

Here (i, j) denotes the user-item pair sampled from the unknown interaction distribution $\mathcal{D}$, θ_i, $\theta_j \in \mathbb{R}^d$ (the i, j-th row of Θ) denotes their embedding vectors respectively, and the loss $\ell(\theta_i, \theta_j; y_{ij})$ denotes the prediction loss for their interaction $y_{ij} \in \{-1, +1\}$ (e.g., logistic loss). We further let

$$p_i = \sum_{j\in V} D(i,j), \;\; \forall i \in U; \;\; p_j = \sum_{i\in U} D(i,j), \;\; \forall j \in V, \tag{2}$$

denote the marginal distribution over U and V.

Remark 2.1. Our analysis also allows treatment of additional network structure (with parameters denoted by $\mathcal{W}$) that takes nonlinear transformation of embedding vectors, e.g., $f(\Theta, \mathcal{W}) = \mathbb{E}_{(i,j)\sim\mathcal{D}}\ell(\theta_i, \theta_j, \mathcal{W}; y_{ij})$. We omit their explicit treatment for presentation simplicity. In addition, although we mainly discuss in the context of recommendation, our analysis and results only relies on sparsity of stochastic gradient and the imbalancedness of token distributions, which allow one to extend our results to other embedding learning problems (e.g., language model pretraining).

The full algorithmic descriptions of our proposed Frequency-aware Stochastic Gradient Descent (FA-SGD) algorithm are presented in Algorithm 1. Note that randomly outputting a historical iterate is commonly adopted in literature for showing convergence of stochastic gradient descent type algorithms for non-convex problems (Ghadimi & Lan, 2013). In practice, we can simply use the last iterate Θ^T as the output solution. In addition, Section 3.3 presents CF-SGD (Algorithm 2), which does not need the token distribution as the input and can estimate it in an online fashion.

At iteration t, FA-SGD samples $(i_t, j_t) \sim \mathcal{D}$, and obtain the sparse stochastic gradient g^t defined in (3). Note that only the i_t-th and j_t-th row of g_t are non-zero. One can readily verify that $\mathbb{E}_{(i_t,j_t)\sim\mathcal{D}}\left[g_t\right] = \nabla_\Theta f(\Theta^t)$. Going forward, we will denote ∇f_k^t as the k-th row of gradient $\nabla f(\Theta^t)$, and g_k^t as the k-th row of stochastic gradient g_t. Note that we have

$$g_t = \begin{bmatrix} \mathbf{0}^\top \\ \nabla_{\theta_{i_t}} \ell(\theta_{i_t}, \theta_{j_t}; y_{i_t,j_t})^\top \\ \vdots \\ \nabla_{\theta_{j_t}} \ell(\theta_{i_t}, \theta_{j_t}; y_{i_t,j_t})^\top \\ \mathbf{0}^\top \end{bmatrix} \tag{3}$$

$$\mathbb{E}_{j_t}\left[g_{i_t}^t | i_t = i\right] = \nabla f_i^t / p_i, \;\; \mathbb{E}_{i_t}\left[g_{j_t}^t | j_t = j\right] = \nabla f_j^t / p_j. \tag{4}$$

We further denote $\delta_k^t = \frac{1}{p_k} f_k^t - g_k^t$ for all $k \in X$. Then by definition $\mathbb{E}\left[\delta_{i_t}^t | i_t = i\right] = 0$ and $\mathbb{E}\left[\delta_{j_t}^t | j_t = j\right] = 0$ for all $i \in U, j \in V$. We pose the following assumptions on the its variance.

Assumption 1 (Bounded conditional variance)**.** *We assume that the variance of $\delta_{i_t}^t$ is bounded. That is, there exists $\{\sigma_k^2\}_{k \in X}$, such that*

$$\mathbb{E}\left[\|\delta_{i_t}\|^2 | i_t = i\right] \leq \sigma_i^2, \ \mathbb{E}\left[\|\delta_{j_t}\|^2 | j_t = j\right] \leq \sigma_j^2, \quad \forall i \in U, j \in V. \tag{5}$$

Assumption 1 allows us to provide a finer characterization on the variance of stochastic gradient compared to typical variance assumption in literature. To illustrate, recall that standard assumption in the stochastic optimization literature assumes $\mathrm{Var}(g^t) = \mathbb{E}\|\nabla_\Theta f^t - g^t\|^2 \leq \sigma^2$ for some universal constant $\sigma > 0$. Consider an extreme setting, where we have exact gradient for the sampled user-item pair, i.e., $g_{i_t}^i = \frac{1}{p_{i_t}} \nabla f_{i_t}^t$ and $g_{j_t}^i = \frac{1}{p_{j_t}} \nabla f_{j_t}^t$, then we have $\sigma_k = 0$ for all $k \in X$. In contrast, the variance of g^t is still non-zero. In general setting, we can bound the variance as shown in the following proposition. Note that the variance lower bound arises naturally form the extreme sparsity of the stochastic gradient.

Proposition 2.1. *Given Assumption 1, we have*

$$\sum_{k \in X} (1/p_k - 1) \left\|\nabla f_k^t\right\|^2 \leq \mathrm{Var}(g^t) \leq \sum_{k \in X} p_k \sigma_k^2 + \sum_{k \in X} (1/p_k - 1) \left\|\nabla f_k^t\right\|^2. \tag{6}$$

Assumption 2 (Smoothness of prediction loss)**.** *We assume $\ell(u, v; y)$ is symmetric w.r.t. u and v for any $y \in \{-1, +1\}$, and there exists $L > 0$ such that $\left\|\nabla_{uu}^2 \ell(\cdot, \cdot; \cdot)\right\|_2 \leq L$, $\left\|\nabla_{uv}^2 \ell(\cdot, \cdot; \cdot)\right\|_2 \leq L$.*

The assumption on the symmetry of ℓ is readily satisfied by almost all neural network architecture. In essence, this assumption only requires that the parameterization of embedding vector is token agnostic. On the other hand, the spectral upper bound on the Hessian matrix is a standard assumption in optimization literature.

3 Theoretical Results

We first present the convergence results of FA-SGD and standard SGD for embedding learning problem formulated in (1), and discuss the advantage that FA-SGD offers when the token distribution $\{p_k\}_{k \in X}$ is highly imbalanced. We further propose a variant, named CF-SGD, which can estimate frequency information in an online fashion and still provably enjoys the benefits of FA-SGD.

3.1 Convergence of FA-SGD and standard SGD

Theorem 3.1 (FA-SGD)**.** *With Assumption 1 and 2, take learning rate policy to be*

$$\eta_k^t = \min\left\{1/(4L), \alpha/\sqrt{T p_k}\right\}, \tag{7}$$

where T denotes the total number of iterations, and $\alpha = \sqrt{\left(f(\Theta^0) - f^*\right) / \left(L \sum_{l \in X} p_l \sigma_l^2\right)}$, we have

$$\mathbb{E}\|\nabla f_k^\tau\|^2 = \mathcal{O}\left(\frac{L\left(f(\Theta^0) - f^*\right)}{T} + \frac{\sqrt{p_k}\sqrt{\sum_{l \in X} p_l \sigma_l^2 (f(\Theta^0) - f^*) L}}{\sqrt{T}}\right), \quad \forall k \in X. \tag{8}$$

Remark 3.1 (Connection with Stochastic Block Coordinate Descent). Our FA-SGD shares some similarities with Stochastic Block Coordinate Descent (SBCD) (Nesterov, 2012; Dang & Lan, 2015; Richtárik & Takáč, 2014) applied to problem (1), in the sense that each iteration we sample certain blocks of variables ($\theta_{i_t}, \theta_{j_t}$ in our case), and only update the sampled blocks by following its stochastic gradient. Different from SBCD, the stochastic gradient of the block variable $g_{i_t}^t$ in the FA-SGD is *biased*, as shown in (4). Note that with unbiased stochastic gradient, SBCD method typically converges slower than standard SGD by a factor that can be as large as number of blocks. As a concrete example, when the token distribution is uniform, SBCD converges slower than standard SGD by a factor of $|X|$, hence slower than FA-SGD by a factor of $|X|$ from Corollary 3.1 developed later.

Recall that from Proposition 2.1, the variance of stochastic gradient is heavily influenced by the population gradient $\nabla_\Theta f(\Theta)$, and can be huge whenever the population gradient is, presumably in the early phase of training. This relationship is also supported by empirical findings in Zhang et al. (2019) (Figure 2a), where the authors show that for BERT pretraining, the noise distribution in stochastic gradient g^t is highly non-stationary, which has large variance in the beginning of the training and smaller variance at the end of training. Since existing analysis of SGD in literature assumes a constant variance bound for the stochastic gradient, our observation in Proposition 2.1 requires an alternative analysis of SGD for problem (1).

To obtain the convergence rate of standard SGD in the presence of iterate-dependent variance (6), our key insight is to tailor the convergence analysis to the sparsity of the stochastic gradient for problem (1). We show the convergence of standard SGD as the following.

Theorem 3.2 (Standard SGD)**.** *With Assumption 1 and 2, 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*

$$\mathbb{E}\left\|\nabla f_k^\tau\right\|^2 = \mathcal{O}\left(\frac{L\left(f(\Theta^0)-f^*\right)}{T} + \frac{\sqrt{\sum_{l\in X} p_l^2\sigma_l^2(f(\Theta^0)-f^*)L}}{\sqrt{T}}\right), \quad \forall k \in X. \tag{9}$$

Note that both FA-SGD and standard SGD attain a rate of $\mathcal{O}(1/\sqrt{T})$. Compared to existing rates of standard SGD (Ghadimi & Lan, 2013), we do not require constant variance bound on stochastic gradient, as we have discussed above. Compared to existing rates of adaptive learning rate algorithms (Zhou et al., 2018a; Chen et al., 2018), both rates obtained here exhibits *dimension-free* property. We emphasize here that due to the dimension-free nature of the bounds for both SGD and FA-SGD, *we do not claim the proposed FA-SGD has better dependence on dimension, which is the main motivation of adaptive algorithms* (Duchi et al., 2011; Kingma & Ba, 2014; Reddi et al., 2019). Instead, the major difference on the convergence of FA-SGD (8) and that of standard SGD (9) is that the former one is *token-dependent*. Specifically, for FA-SGD, each token $k \in X$ has its own convergence characterization, while all the tokens have the same convergence characterization in the standard SGD. We first make a simple observation stating the equivalence of FA-SGD and standard SGD, when the token distribution $\{p_k\}_{k\in X}$ is uniform.

Corollary 3.1 (Uniform Distribution)**.** *Suppose the user distribution* $\{p_i\}_{i\in U}$ *and item distribution* $\{p_j\}_{j\in V}$ *is the uniform distribution. Then FA-SGD and standard SGD is equivalent to each other, in terms of both algorithmic execution and convergence rate.*

3.2 When does FA-SGD outperform standard SGD?

We show FA-SGD shines when the token distribution $\{p_k\}_{k\in X}$, defined in (2), is highly imbalanced. Before we present detailed discussions, we make an important remark that highly imbalanced token distributions are ubiquitous in social systems, presented in the form *power-law*. Examples of such distributions include the degree of individuals in the social network (Muchnik et al., 2013); the frequency of words in natural language (Zipf, 2016); citations for academic papers (Brzezinski, 2015); number of links on the internet (Albert et al., 1999). For more discussions on power-law distributions in social and natural systems, we refer readers to Kumamoto & Kamihigashi (2018).

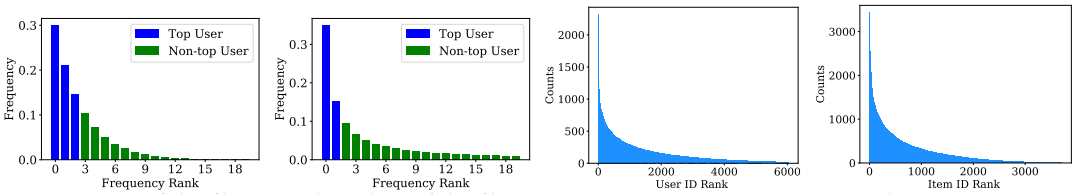
In Figure 1c, 1d we plot the user and item counting distribution of Movielens-1M dataset. One could clearly see that the user and item distributions are highly imbalanced, with a small percentages of users/items taking up the majority of rating records. We defer details on the skewness of token distributions for Criteo dataset to Appendix C.

To illustrate the comparative advantage of FA-SGD when the token distribution $\{p_k\}_{k\in X}$ is highly skewed. We consider two classes of distribution families with different tail properties, one with exponential tail, and one with polynomial tail.

Corollary 3.2 (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*

$$p_{i_n} \propto \exp(-\tau n), \;\; p_{j_m} \propto \exp(-\tau m), \quad \forall n \in [|U|], m \in [|V|] \tag{10}$$

for some $\tau > 0$. *Define* U_T *as the set of users whose frequencies are within* e*-factor from the highest frequency:* $U_T = \{i_n : n \leq \frac{1}{\tau}\}$, *and* V_T *similarly as* $V_T = \{j_m : m \leq \frac{1}{\tau}\}$. *We refer to* U_T *as the top users, and* V_T *as the top items.*



(a) Exponential Tail (b) Polynomial Tail (c) User counts (d) Item counts

Figure 1: Token distribution with an exponential and polynomial tail, and the user/item counting distributions for Movielens-1M dataset.

Then given $|U|, |V| \geq \frac{1}{\tau}$, the proposed 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) $\mathbb{E}\left\|\nabla f_{i_n}^{\tau}\right\|^2$ 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) $\mathbb{E}\left\|\nabla f_{j_m}^{\tau}\right\|^2$ can converge faster by a factor of $\Omega\left\{\exp\left(\tau(m - |V_T|)\right)\right\}$ for $j_m \in V \setminus V_T$.

We remark that $|U|, |V| \geq \frac{1}{\tau}$ is a very mild condition, as it only requires that the most infrequent user/item should have its frequency smaller than the most frequent user/item by at least a factor of e. i.e., the non-top user/item set $U \setminus U_T$, $V \setminus V_T$ is nonempty, This is readily satisfied by the token distributions in recommendation systems and natural language modeling (Celma, 2010; Zipf, 2016), where the lowest frequency is at least orders of magnitude smaller than the highest frequency. The factor of e in defining U_T, V_T can also be readily replaced by any constant larger than 1.

From Corollary 3.2, we can see that FA-SGD improves significantly over standard SGD for user/item distribution with exponential tail. Specifically, FA-SGD achieves the same convergence rate of top users/items compared to SGD, meanwhile it significantly improves the convergence of the non-top users/items. Moreover, the strength of such an *improvement increases exponentially as we move towards the tail users/items*.

Corollary 3.3 (Polynomial 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*

$$p_{i_n} \propto n^{-\nu}, \;\; p_{j_m} \propto m^{-\nu}, \quad \forall n \in [|U|], m \in [|V|] \tag{11}$$

for some $\nu \geq 2$. Define U_T as the set of users whose frequencies are within 2-factor from the highest frequency: $U_T = \{i_n : n^{-\nu} \geq 1/16\}$, and V_T similarly as $V_T = \{j_m : m^{-\nu} \geq 1/16\}$. We refer to U_T as the top users, and V_T as the top items.

Then given $|U|, |V| \geq 16^{1/\nu}$, the 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) $\mathbb{E}\left\|\nabla f_{i_n}^{\tau}\right\|^2$ can converge faster by a factor of $\Omega\left\{\left(\frac{n}{|U_T|}\right)^{\nu}\right\}$ for each $i_n \in U \setminus U_T$;
(3) $\mathbb{E}\left\|\nabla f_{j_m}^{\tau}\right\|^2$ can converge faster by a factor of $\Omega\left\{\left(\frac{m}{|V_T|}\right)^{\nu}\right\}$ for each $j_m \in V \setminus V_T$.

We remark that polynomial tail (37) is also the prototypical example of the power law distribution class for modeling social behaviors (Kumamoto & Kamihigashi, 2018). The constant 2 in the condition $\nu \geq 2$ can be replaced by any constant strictly larger than 1, with slight changes to the constant factor in the statements of the corollary.

From Corollary 3.3, we can see that FA-SGD improves significantly over standard SGD for user/item distribution with polynomial tail. Specifically, FA-SGD achieves the same convergence rate of top users/items compared to SGD, meanwhile it significantly improves the convergence of the non-top users/items. Moreover, the strength of such an *improvement increases in polynomial order as we move towards the tail users/items*.

3.3 Online Estimation of Frequency Information

In certain application scenarios, the token distribution $\{p_k\}_{k \in X}$ can be unknown in advance of learning. To apply FA-SGD, one needs to employ a preprocessing step in order to estimate the token distribution to a high accuracy, and then run the algorithm with estimated token distribution. Such a preprocessing step often requires additional human efforts and data. To remove such an undesirable

Algorithm 2 Counter-based Frequency-aware Stochastic Gradient Descent

Input: Total iteration number T.
Initialize: $\Theta^0 \in \mathbb{R}^{N\times d}$, counter sample $\tau \sim \mathrm{Unif}(\{T/2, \dots, T\})$.
for $t = 0, \dots \tau$ **do**
(1) Sample $(i_t, j_t) \sim \mathcal{D}$, calculate the stochastic gradient

$$g^t_{i_t} = \nabla_{\theta_{i_t}} \ell(\theta_{i_t}, \theta_{j_t}; y_{i_t,j_t}), \;\; g^t_{j_t} = \nabla_{\theta_{j_t}} \ell(\theta_{i_t}, \theta_{j_t}; y_{i_t,j_t})$$

(2) Compute counter-based learning rate $\widehat{\eta}^t_{i_t}\left(c^t_{i_t}\right), \widehat{\eta}^t_{j_t}\left(c^t_{j_t}\right)$ specified by (12)
(3) Update parameters

$$\theta^{t+1}_{i_t} = \theta^t_{i_t} - \widehat{\eta}^t_{i_t} g^t_{i_t}, \;\; \theta^{t+1}_i = \theta^t_i, \;\; \forall i \in U, i \neq i_t$$
$$\theta^{t+1}_{j_t} = \theta^t_{j_t} - \widehat{\eta}^t_{j_t} g^t_{j_t}, \;\; \theta^{t+1}_j = \theta^t_j, \;\; \forall j \in V, j \neq j_t$$

(4) Update counters

$$c^{t+1}_{i_t} = c^t_{i_t} + 1, \;\; c^{t+1}_i = c^t_i, \;\; \forall i \in U, i \neq i_t$$
$$c^{t+1}_{j_t} = c^t_{j_t} + 1, \;\; c^{t+1}_j = c^t_j, \;\; \forall j \in V, j \neq j_t$$

end for
Output: Θ^τ

preprocessing step, below we present an online variant of FA-SGD, which uses the counter of tokens collected during training to estimate the token distribution dynamically. We show that the proposed **C**ounter-based **F**requency-aware **S**tochastic **G**radient **D**escent (CF-SGD) is able to retain the benefits of FA-SGD despite unknown token distribution.

Theorem 3.3 (Counter-based FA-SGD)**.** *In addition to Assumption 1 and 2, suppose $\|\nabla f(\cdot)\| \leq G$. Take counter-based learning rate policy in Algorithm 2 to be*

$$\widehat{\eta}^t_k(c^t_k) = \min\left\{1/(4L), 1/\sqrt{T\widehat{p}^t_k}\right\}, \quad \widehat{p}^t_k = c^t_k/t, \quad \forall k \in X, t \in [T], \tag{12}$$

where T denotes the total number of iterations, $\alpha = \sqrt{M_f/\left(L\sum_{l\in X} p_l\sigma_l^2\right)}$ and $M_f = f(\Theta^0) - f^* + \sum_{k\in X} p_k\sigma_k^2/L$, we have

$$\mathbb{E}\left\|\nabla f^\tau_k\right\|^2 = \mathcal{O}\left(\frac{LM_f}{T} + \frac{\sqrt{p_k}\sqrt{\sum_{l\in X} p_l\sigma_l^2 L\left(f(\Theta^0) - f^*\right)}}{\sqrt{T}} + \frac{\sqrt{p_k}\left(\sum_{l\in X} p_l\sigma_l^2\right)}{\sqrt{T}}\right), \forall k \in X. \tag{13}$$

for $T \geq \max\left\{\min_{l\in X}\frac{1}{p_l}, \frac{2\log G - \log\left(M_f(1/2L + \alpha/\sqrt{p_k})\right)}{p_k}\right\}$.

We believe the assumption on gradient bound $\|\nabla f(\cdot)\| \leq G$ is not strictly necessary and can be removed with more refined analysis. Nevertheless, the requirement on T only logarithmically depends on the gradient bound G. In addition, we highlight that the convergence characterization in Theorem 3.3 is still token-dependent. Specifically, we can show that despite not knowing token distribution beforehand, CF-SGD can gain the same advantages that FA-SGD enjoys over SGD.

Corollary 3.4 (Exponential Tail)**.** *Suppose we have the same set of conditions given in Corollary 3.2, and $\sigma/\sqrt{L\left(f(\Theta^0) - f^*\right)} \leq 1$. Define U_T as the set of users whose frequencies are within e-factor from the highest frequency: $U_T = \{i_n : n \leq \frac{1}{\tau}\}$, and V_T similarly as $V_T = \{j_m : m \leq \frac{1}{\tau}\}$. We refer to U_T as the top users, and V_T as the top items.*

Then given $|U|, |V| \geq \frac{1}{\tau}$, the proposed CF-SGD, compared to standard SGD:

(1) Obtains the same rate of convergence, for the top users U_T and top items V_T;
(2) $\mathbb{E}\left\|\nabla f^\tau_{i_n}\right\|^2$ 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) $\mathbb{E}\left\|\nabla f^\tau_{j_m}\right\|^2$ can converge faster by a factor of $\Omega\left\{\exp\left(\tau(m - |V_T|)\right)\right\}$ for $j_m \in V \setminus V_T$.

Corollary 3.5 (Polynomial Tail)**.** *Suppose we have the same set of conditions given in Corollary 3.3, and $\sigma/\sqrt{L\left(f(\Theta^0) - f^*\right)} \leq 1$. Define U_T as the set of users whose frequencies are within*

2-factor from the highest frequency: $U_T = \{i_n : n^{-\nu} \geq 1/16\}$, and V_T similarly as $V_T = \{j_m : m^{-\nu} \geq 1/16\}$. We refer to U_T as the top users, and V_T as the top items.

Then given $|U|, |V| \geq 16^{1/\nu}$, the 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) $\mathbb{E}\left\|\nabla f_{i_n}^{\tau}\right\|^2$ can converge faster by a factor of $\Omega\left\{\left(\frac{n}{|U_T|}\right)^{\nu}\right\}$ for each $i_n \in U \setminus U_T$;
(3) $\mathbb{E}\left\|\nabla f_{j_m}^{\tau}\right\|^2$ can converge faster by a factor of $\Omega\left\{\left(\frac{m}{|V_T|}\right)^{\nu}\right\}$ for each $j_m \in V \setminus V_T$.

The proofs of Corollary 3.4 and 3.5 follow similar lines as in the proofs of Corollary 3.2 and 3.3, which we defer to Appendix D

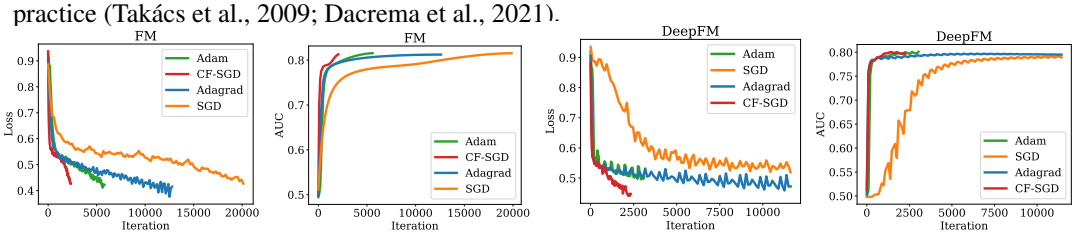
4 Experiments

We conduct extensive experiments to verify the effectiveness of our proposed algorithms and our developed theories, on both publicly available benchmark recommendation datasets, and a large-scale industrial recommendation system. Additional experiments on learning Word2Vec embeddings (Mikolov et al., 2013a) are presented in Appendix B, demonstrating the general applicability of the FA/CF-SGD. We list key elements of our experiment setup for benchmark datasets below.

- **Datasets**: Benchmark recommendation datasets MovieLens-1M[1] and Criteo[2].
- **Models**: Factorization Machine (FM) (Rendle, 2010), and DeepFM (Guo et al., 2017).
- **Metric**: Training loss (cross-entropy loss), and test AUC (Area Under the ROC Curve).
- **Baseline algorithms**: SGD, Adam (Kingma & Ba, 2014), Adagrad (Duchi et al., 2011). Note that Adam and Adagrad are widely popular in training recommendation systems and language models.

We also empirically verify that the token distributions for both Movielens-1M (Figure 1) and Criteo (Appendix C) dataset are highly imbalanced, with most of the token distributions having a clear polynomially or exponentially decaying tail.

Since CF-SGD does not require frequency information, which is a huge practical benefit compared to FA-SGD, in our experiments we mainly evaluate our proposed CF-SGD against the baseline algorithms. To ensure a fair comparison, for each dataset and model type, we carefully tune the learning rate of each algorithm for best performance[3]. We apply early stopping and stop training whenever the validation AUC do not increase for 2 consecutive epochs, which is widely adopted in practice (Takács et al., 2009; Dacrema et al., 2021).



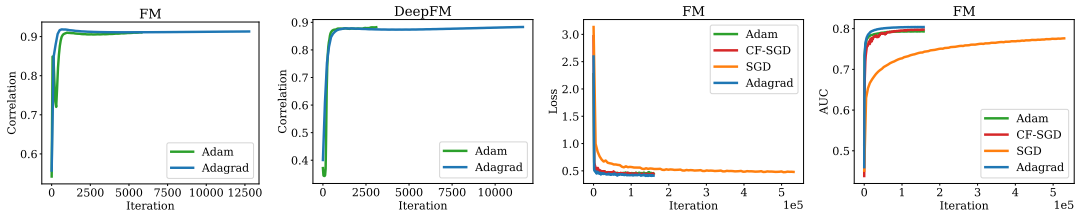
(a) Training loss (b) Validation AUC (c) Training loss (d) Validation AUC

Figure 2: Movielens-1M dataset with FM and DeepFM model. CF-SGD significantly outperforms standard SGD, and is highly competitive against Adam, Adagrad.

Movielens-1M: We can observe from Figure 2 that for FM and DeepFM model: (1) SGD yields the slowest convergence in training loss and AUC. (2) The proposed CF-SGD yields significantly faster convergence than SGD for training loss. In addition, CF-SGD converges even faster than the adaptive learning algorithms in the early stage of training; (3) All the algorithms eventually reaches peak AUC around 81.0%, while CF-SGD attains the peak AUC much faster than baseline algorithms. These empirical observations help us confirm the effectiveness of the proposed CF-SGD algorithm. We further make an empirical observation that draws a close connection between adaptive algorithms and CF-SGD. We plot the second-order gradient moment maintained by Adagrad and Adam against the estimated frequency maintained by CF-SGD. Surprisingly, the second-order gradient moment quickly develops a close-to linear relationship with the frequency information accumulated by CF-SGD (Figure 3a,3b) . This observation suggests that Adagrad and Adam are exploiting frequency information implicitly to a large extent.

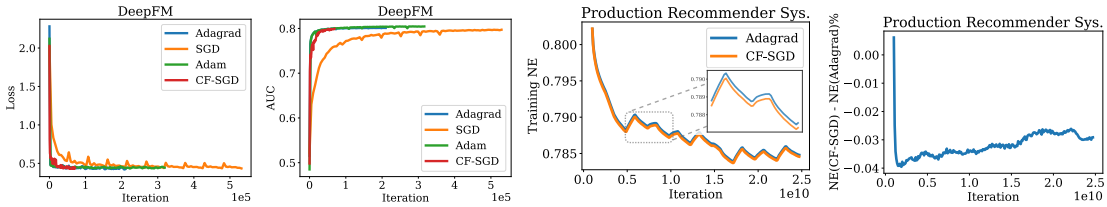
[1] https://grouplens.org/datasets/movielens/1m/
[2] https://ailab.criteo.com/ressources/
[3] Further details on architecture and hyper-parameter choice can be found at Appendix A.

Criteo: We observe qualitative behavior of CF-SGD similar to Movielens-1M dataset, as can be seen in Figure 3c,3d, 4a,4b.



(a) Correlation (b) Correlation (c) Training loss (d) Validation AUC

Figure 3: (a-b) Second-order gradient moment correlates linearly with frequency maintained by CF-SGD; (c-d) Comparisons on Criteo dataset with FM model.



(a) Training loss (b) Validation AUC (c) Train NE Curve (d) Train NE Diff %

Figure 4: (a-b) Comparisons on Criteo dataset with DeepFM model; (c-d) Comparisons on a industrial-scale recommendation dataset with an ultra-large recommender model.

Industrial Recommendation System: We train an ultra-large industrial recommendation model with the proposed CF-SGD. The training data contains 10 days of user-item interaction records, with $\sim$2.5 billion examples per day (25 billion examples in total). We use around 800 features, with $\sim$100 million average number of tokens per feature. We compare CF-SGD with Adagrad, which has been carefully tuned in production usage. For both algorithms, we use a batch size of 64k and do one-pass training. Different from benchmark academic datasets, we use Normalized Entropy (NE) as the evaluating metric (He et al., 2014) (smaller is better), which is the cross-entropy loss normalized by the entropy of background click through rate. Note that due to numerous iterations of the production model, any relative improvement $\sim 0.02\%$ is considered to be significant. In Figure 4c, 4d we compare the training NE curve CF-SGD and Adagrad, we can see that CF-SGD shows faster convergence than Adagrad (see NE difference % in Figure 4d). From Table 1 we can observe that CF-SGD also improves over Adagrad during the serving phase.

Alg	NE	Diff %
Adagrad	0.78643	0.0
CF-SGD	0.78628	-0.02

Table 1: Eval NE Diff %

Memory Efficiency: On top of the above empirical evidences showing that CF-SGD learns *fast* – faster than standard SGD, and comparable (if not better) to adaptive algorithms, we further highlight that CF-SGD learns *cheap*. Specifically, adaptive algorithms require additional memory to store history information for each parameter. For an embedding table of size $N \times d$ (N tokens, d being embedding dimension), the memory needed is at least $3N \times d$ for Adam (first/second-order gradient moment), and $2N \times d$ for Adagrad (second-order gradient moment). In sharp contrast, CF-SGD only requires N additional memory, for storing the estimated frequency. Since the choice of typical embedding dimension d exceeds 64 (Yin & Shen, 2018), adaptive algorithms require memory at least twice the size of the embedding table, while CF-SGD requires negligible memory overhead. Note the industrial recommendation model in our experiments has a size over multiple terabytes, with above 95% of consumed by embedding tables. Doubling the memory footprint by using standard Adam/Adagrad is infeasible in terms of both engineering and environmental concern.

5 Conclusion

We propose (Counter-based) Frequency-aware SGD for embedding learning problems, which adopts frequency-dependent learning rate schedule for each token. We demonstrate provable benefits that FA/CF-SGD enjoy over standard SGD for imbalanced token distributions, with extensive experiments supporting our theoretical findings. Our empirical findings also suggest that adaptive algorithms can implicitly exploit frequency information and hence share close connections with the proposed algorithms, this connection might be helpful in the direct analysis of adaptive algorithms for embedding learning problems, which we leave as a future direction.

REFERENCES

Réka Albert, Hawoong Jeong, and Albert-László Barabási. Diameter of the world-wide web. *nature*, 401(6749):130–131, 1999.

Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *arXiv preprint arXiv:1912.02365*, 2019.

Michal Brzezinski. Power laws in citation distributions: evidence from scopus. *Scientometrics*, 103 (1):213–228, 2015.

Òscar Celma. The long tail in recommender systems. In *Music Recommendation and Discovery*, pp. 87–107. Springer, 2010.

Xiangyi Chen, Sijia Liu, Ruoyu Sun, and Mingyi Hong. On the convergence of a class of adam-type algorithms for non-convex optimization. *arXiv preprint arXiv:1808.02941*, 2018.

Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

Maurizio Ferrari Dacrema, Simone Boglio, Paolo Cremonesi, and Dietmar Jannach. A troubling analysis of reproducibility and progress in recommender systems research. *ACM Transactions on Information Systems (TOIS)*, 39(2):1–49, 2021.

Cong D Dang and Guanghui Lan. Stochastic block mirror descent methods for nonsmooth and stochastic optimization. *SIAM Journal on Optimization*, 25(2):856–881, 2015.

Alexandre Défossez, Léon Bottou, Francis Bach, and Nicolas Usunier. A simple convergence proof of adam and adagrad. *arXiv preprint arXiv:2003.02395*, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Yoel Drori and Ohad Shamir. The complexity of finding stationary points with stochastic gradient descent. In *International Conference on Machine Learning*, pp. 2658–2667. PMLR, 2020.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. Deepfm: a factorization-machine based neural network for ctr prediction. *arXiv preprint arXiv:1703.04247*, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Xinran He, Junfeng Pan, Ou Jin, Tianbing Xu, Bo Liu, Tao Xu, Yanxin Shi, Antoine Atallah, Ralf Herbrich, Stuart Bowers, et al. Practical lessons from predicting clicks on ads at facebook. In *Proceedings of the Eighth International Workshop on Data Mining for Online Advertising*, pp. 1–9, 2014.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Shin-Ichiro Kumamoto and Takashi Kamihigashi. Power laws in stochastic processes for social phenomena: An introductory review. *Frontiers in Physics*, 6:20, 2018.

Liyuan Liu, Xiaodong Liu, Jianfeng Gao, Weizhu Chen, and Jiawei Han. Understanding the difficulty of training transformers. *arXiv preprint arXiv:2004.08249*, 2020.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013a.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119, 2013b.

Lev Muchnik, Sen Pei, Lucas C Parra, Saulo DS Reis, José S Andrade Jr, Shlomo Havlin, and Hernán A Makse. Origins of power-law degree distribution in the heterogeneity of human activity in social networks. *Scientific reports*, 3(1):1–8, 2013.

Yu Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL https://aclanthology.org/D14-1162.

Steven T Piantadosi. Zipf?s word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, 21(5):1112–1130, 2014.

S Reddi, Manzil Zaheer, Devendra Sachan, Satyen Kale, and Sanjiv Kumar. Adaptive methods for nonconvex optimization. In *Proceeding of 32nd Conference on Neural Information Processing Systems (NIPS 2018)*, 2018.

Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237*, 2019.

Steffen Rendle. Factorization machines. In *2010 IEEE International conference on data mining*, pp. 995–1000. IEEE, 2010.

Peter Richtárik and Martin Takáč. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144(1):1–38, 2014.

Joaquim Santos, Bernardo Consoli, and Renata Vieira. Word embedding evaluation in downstream tasks and semantic analogies. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 4828–4834, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.594.

Gábor Takács, István Pilászy, Bottyán Németh, and Domonkos Tikk. Scalable collaborative filtering approaches for large recommender systems. *The Journal of Machine Learning Research*, 10:623–656, 2009.

Rachel Ward, Xiaoxia Wu, and Leon Bottou. Adagrad stepsizes: Sharp convergence over nonconvex landscapes, from any initialization. *arXiv preprint arXiv:1806.01811*, 2, 2018.

Zi Yin and Yuanyuan Shen. On the dimensionality of word embedding. *arXiv preprint arXiv:1812.04224*, 2018.

Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank J Reddi, Sanjiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models? *arXiv preprint arXiv:1912.03194*, 2019.

Dongruo Zhou, Jinghui Chen, Yuan Cao, Yiqi Tang, Ziyan Yang, and Quanquan Gu. On the convergence of adaptive gradient methods for nonconvex optimization. *arXiv preprint arXiv:1808.05671*, 2018a.

Guorui Zhou, Xiaoqiang Zhu, Chenru Song, Ying Fan, Han Zhu, Xiao Ma, Yanghui Yan, Junqi Jin, Han Li, and Kun Gai. Deep interest network for click-through rate prediction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1059–1068, 2018b.

George Kingsley Zipf. *Human behavior and the principle of least effort: An introduction to human ecology*. Ravenio Books, 2016.