LSH Microbatches for Stochastic Gradients: Value in Rearrangement

Eliav Buchnik*  
Tel Aviv University  
Google Research, Israel  
eliavbuh@gmail.com

Edith Cohen*  
Google Research USA  
Tel Aviv University  
edith@cohenwang.com

Avinatan Hassidim  
Google Research, Israel

Abstract

Metric embeddings are immensely useful representations of associations between entities (images, users, search queries, words, and more). Embeddings are learned by optimizing a loss objective of the general form of a sum over example associations. Typically, the optimization uses stochastic gradient updates over minibatches of examples that are arranged independently at random. In this work, we propose principled methods of structuring the arrangements of training examples aimed at accelerating the training. Our arrangements consist of randomized microbatches of examples from a distribution that is guided by the structure of the example associations but respects a specified marginal distribution of training examples. We present efficient microbatch generators and experimentally demonstrate training time acceleration of 3-20%. Structured arrangements emerge as a powerful and novel performance knob for SGD that is independent and complementary to other SGD hyperparameters and thus is a candidate for wide deployment.

1 Introduction

Metric embeddings of entities that are trained to capture example associations are common representations that also allow for inference of associations not present in the data. Embeddings are used in complex learning tasks or directly applied for similarity and recommendations tasks. Example usage domains includes embeddings of text document from occurrences of words [3, 10, 8], users and videos from watch or ratings [16], words from co-occurrence frequencies in a corpus [17], and nodes in a graph from co-occurrence in short random walks [20]. The example associations may involve entities of the same type (word co-occurrences, video co-watch, social) or different types (such as users and products) and often are distilled by reweighing frequencies of raw interactions [23, 8, 17, 19].

Embeddings are computed by minimizing a loss objective of the form of a sum over example associations. The optimization starts with random initialization followed by gradient updates. In modern applications, the objective can have billions of terms or more, and the de facto method at such scale is stochastic gradient descent (SGD) [21, 15, 22, 12, 17]. The terms (examples) are randomly grouped into minibatches. Gradient updates, that are equal in expectation to the full gradient, are then computed sequentially for each minibatch. SGD is much more efficient than working with full gradients and the minibatch size determines the amount of concurrency. There are numerous tunable parameters and methods aimed to improve quality and speed of convergence. Some notable recent work includes per-parameter tuning of the learning rate [9] and altering the distribution of training examples by gradient magnitude [2], negatives selection with triplet loss [24], clustering [11] and diversity criteria [26].

In this work we introduce principled schemes that control the arrangement of examples into minibatches. Note that arrangement tuning is separate and orthogonal to optimizations knobs that alter the distribution of training examples, learning rate, or minibatch size. The baseline practice of independent arrangements places examples into minibatches
independently at random. This practice is supported by classic SGD convergence analysis and has the upside of controlling the variance of the stochastic gradients. We make a novel case here for the antithesis of coordinated arrangements, where corresponding associations are much more likely to be included in the same minibatch. We show that coordination offers different upsides: At the micro level, updates are more effective in pulling vectors of similar entities closer. At a macro level, the examples in small fractions of epochs encode (in expectation) the similarity structure in the full set of example associations whereas independent arrangement disperse that information. This “self similarity” of the training sequence effectively allows a single epoch to act as multiple passes.

We specify coordinated arrangements through a distribution on randomized subsets of associations which we refer to as microbatches. Basic coordinated microbatches co-place corresponding associations to the maximum extent possible while adhering to the marginal distribution of training examples. Locality Sensitive Hashing (LSH) maps allow for refining our microbatches so that corresponding associations are more likely to be co-placed when the overall similarity of the entities is higher. The LSH maps we apply leverage some available coarse proxy of entity similarity. A readily available first-order proxy is the similarity of the sparse association vectors. Another proxy is an embedding obtained by a weaker model. We present efficient generators of basic and refined microbatches for both LSH functions. Finally, microbatches are independently grouped into minibatches of desired size, which allows us to retain the traditional advantages of independent arrangements at the microbatch level. With this design we are able to tune our microbatches to the problem and stage of training.

We compare the effectiveness of different arrangements through experiments on synthetic stochastic block matrices and on recommendation data sets. The stochastic block data, with its simplicity and symmetry, allows us to factor out the potential effect of other optimizations. We learn that basic coordination is always beneficial earlier in training whereas LSH refinements or even independent arrangements can be more effective later on. A simple tunable method (even with two switch points set as hyperparameters) is often more effective than any one pure method. We obtain consistent 3%-20% reduction in training with respect to the independent arrangements baseline that holds across other training hyperparameters.

The paper is organized as follows. Section 2 presents necessary background on the SGNS loss objective and working with minibatches with one-sided gradient updates. In Section 3 we define LSH microbatches and coordinated minibatch arrangements. In Section 4 we review the generative model for stochastic blocks and the quality measures we apply. In Section 5 we examine properties of coordinated arrangements that are helpful for faster convergence. We report our experiment results comparing different arrangement methods on stochastic blocks and on recommendation data sets in Section 6. We conclude in Section 7.

2 Preliminaries

Our data has the form of associations between a set $F$ of focus entities and a set $C$ of context entities. We use $\kappa_{ij}$ as the association strength between focus $i$ and context $j$. An embedding is a set of vectors $f_i, c_j \in \mathbb{R}^d$ that is trained to minimize a loss objective that encourages $f_i \cdot c_j$ to be larger when $\kappa_{ij}$ is larger. For concreteness, we focus here on Skip Gram with Negative Sampling (SGNS) loss [17]. Examples of positive associations $(i, j)$ are drawn with probability proportional to $\kappa_{ij}$. Random associations are then used as negative examples [14]: Each positive example $(i, j)$ is matched with a set of negative examples of $i$ with random context entities and similarly $j$ with random focus entities. The negative examples provide an antigravity effect that prevents all embeddings to collapse into the same vector. We use the notation $n_{ij}$ for respective weights of negative associations.

The SGNS objective is designed to maximize the log likelihood of the observed associations. The probability of positive and negative associations are respectively modeled using

$$p_{ij} = \sigma(f_i \cdot c_j) = \frac{1}{1 + \exp(-f_i \cdot c_j)} \quad \text{and} \quad 1 - p_{ij} = \sigma(-f_i \cdot c_j) = \frac{1}{1 + \exp(f_i \cdot c_j)}.$$

The likelihood function, which we seek to maximize, can then be expressed as $\Pi_{ij} p_{ij}^{n_{ij}} \Pi_{ij} (1 - p_{ij})^{n_{ij}}$. We equivalently can minimize the negated log likelihood that turns the objective into a sum:

$$L := \sum_{ij} \kappa_{ij} \log p_{ij} + \sum_{ij} n_{ij} \log(1 - p_{ij}).$$
The optimization is performed by random initialization of the embedding vectors followed by stochastic gradient updates. The stochastic gradients are computed for minibatches of examples that include \(b\) positive examples, where \((i, j)\) appears with frequency \(\kappa_{ij}/|\kappa|_1\) and corresponding sets of negative examples.

### 2.1 One-sided updates

We work with one-sided updates, where each minibatch updates only its focus or only its context embedding vectors, and accordingly say that minibatches are designated for focus or context updates. One-sided updates are related to alternating minimization [7] and to decomposition-coordination approaches [5]. For our purposes, one-sided updates facilitate our coordinated arrangements and also allow more precise matching of corresponding sets of negative examples to positive ones. In a focus-updating minibatch, we will generate a random set of \(\lambda\) context vectors \(C'\) and for each positive example \((i, j)\) we generate \(\lambda\) negative examples \((i, j')\) for \(j' \in C'\). The focus embedding \(f_i\) is updated to be closer to \(c_j\) but at the same time repealed (in expectation) from \(C'\) context vectors:

\[
\dot{f}_i \leftarrow f_i - \eta \nabla f_i \left( \log \sigma(f_i \cdot c_j) + \sum_{j' \in C'} \log \sigma(-f_i \cdot c_{j'}) \right)
\]

where \(\eta\) is the applicable learning rate. Symmetrically, for a context-updating minibatch we use a random set of focus vectors \(F'\) as our negative examples and for each positive example \((i, j)\) we perform the update \(c_j \leftarrow c_j - \eta \nabla e_j \left( \log \sigma(f_i \cdot c_j) + \sum_{i' \in F'} \log \sigma(-f_{i'} \cdot c_j) \right)\).

### 3 Minibatch arrangement schemes

Minibatch arrangement schemes determine how examples are organized into minibatches of specified size parameter \(b\). At the core of each arrangement scheme is a distribution \(\mu B\) over subsets of positive examples which we call microbatches. Our minibatch distributions have the property that the marginal probability of each example \((i, j)\) is always equal to \(\kappa_{ij}/|\kappa|_1\). However, subset probabilities vary between schemes and within a scheme we generally will have different distributions \(\mu B_f\) for focus and \(\mu B_c\) for context designations.

Minibatches are obtained from microbatches as specified in Algorithm 1 for focus updates (a symmetric construction applies to context updates). The input is a microbatch distribution \(\mu B_f\), minibatch size parameter \(b\), and a parameter \(\lambda\) that determines the ratio of negative to positive training examples. We draw independent microbatches until we have a total of \(b\) or more positive examples. We then draw \(\lambda\) random contexts \(C'\) and generate \(\lambda\) negative examples \((i, j')\) for \(j' \in C'\) for each positive example \((i, j)\). When training, we alternate between focus and context updating minibatches.

The baseline independent arrangement method (IND) can be placed in this framework using microbatches that consist of a single positive example \((i, j)\) selected with probability \(\kappa_{ij}/|\kappa|_1\) (see Algorithm 2). With coordinated arrangements, the microbatch distribution depends on designation. Algorithm 3 generates basic coordinated microbatches (COO) for focus-updates. These microbatches have the form of a set of positive examples with a shared context.

Basic microbatches have the property that if \(\kappa_{ij} \leq \kappa_{i'j}\) and the positive example \((i, j)\) is included in a basic microbatch then the microbatch would also include the positive example \((i', j)\). It is instinctive to consider the special case of \(\kappa\) with all positive entries being equal: Basic microbatches with focus designation have the form of some context \(j\), and all \((i, j)\) with positive \(\kappa_{ij}\). Our basic microbatches maximize the co-placement probability of two examples with a shared context while respecting the marginal probabilities. A symmetric construction applies to context-update microbatches that maximize the co-placement of two examples with a shared focus. We establish that our microbatch generator respects the marginal probabilities, that is, example \((i, j)\) is included with probability \(\propto \kappa_{ij}\);

---

**Algorithm 1: Minibatch construction (Focus updates)**

*Input: \(\mu B_f, b, \lambda\) // Microbatch distribution, size, negative sampling*

\[
P, N \leftarrow \emptyset
\]

repeat

\[
X \sim \mu B_f; \ P \leftarrow P \cup X
\]

until \(|P| \geq b\)

\[
C' \leftarrow \lambda \text{ contexts selected uniformly at random}
\]

forall example pair \((i, j) \in P\)

\[
\text{forall } j' \in C' \text{ do}
\]

\[
\text{let } N \leftarrow N \cup \{(i, j')\}
\]

return \(P \cup N\)

---
Lemma 3.1. A basic coordinated microbatch (Algorithm 3.1) with focus designation includes a positive example \((i, j)\) with probability \(\kappa_{ij} / \sum_h M_h\), where for context \(h\) we define \(M_h := \max_i \kappa_{ij}\). A context designation basic microbatch includes \((i, j)\) with probability \(\kappa_{ij} / \sum_h N_h\), where \(N_h := \max_j \kappa_{hj}\).

Proof. Consider focus updates (apply a symmetric argument for context updates). For context \(j\), define \(M_j := \max_i \kappa_{ij}\). The probability that \((i, j)\) is selected is the probability that context \(j\) is selected, which is \(M_j / \sum_h M_h\) and then the probability that \(u \leq \kappa_{ij} / M_j\) for \(u \in U[0, 1]\). The total probability is the product of the probabilities of these two events which is \(\kappa_{ij} / \sum_h M_h\).

We preprocess \(\kappa\) and precompute the per-context maxima so that we can efficiently draw a random context with probability proportional to the column maxima. We also generate an index for efficient retrieval, for context \(j\) and \(a\) threshold value \(T\), of all entries \(i\) with \(\kappa_{ij} \geq T\).

3.1 LSH maps

Placement of \((i, j)\) and \((i', j)\) in the same focus updating microbatch results in pulling \(f_i\) and \(f_{i'}\) closer together (see the micro-level property highlighted in Section 3.2). This is helpful when the entities \(i\) and \(i'\) are similar in that they have a close target embedding. Otherwise, the update is anyhow countered by other updates and the placement may have the undesirable effects of increased variance of the stochastic gradients and larger microbatches. In particular, microbatch sizes that exceed the minibatch size parameter mean that we effectively use larger minibatches. This suggests that it would be useful to tune the quality of co-placements so as to decrease unhelpful ones while retaining as many helpful ones as we can. We do this using locality sensitive hashing (LSH) to compute randomized maps of entities to keys. Each map is represented by a vector \(s\) of keys for entities such that similar entities are more likely to obtain the same key. We use these maps to refine our basic microbatches by partitioning them according to keys.

Ideally, our LSH modules would correspond to the similarity captured by the target embedding. However, this creates a chicken-and-egg problem as the target embedding is not available at the start of training and is what we want to compute. Instead, we use LSH modules that are available at the start of training and are only a coarse proxy of the target similarity. We work with two LSH modules based on Jaccard and on Angular LSH. The modules generate maps for either focus or context entities which are applied according to the microbatch designation. We will specify the map generation for focus entities, as maps for context entities can be symmetrically obtained by reversing roles.

Our Jaccard LSH module is outlined in Algorithm 3.2. The probability that two focus entities \(i\) and \(i'\) are mapped to the same key is equal to the weighted Jaccard similarity of their association vectors \(\kappa_i\) and \(\kappa_{i'}\). (For context updates the map is according to the vectors \(\kappa_{i,j}\).)

Lemma 3.2. \([3]\)

\[
\Pr[s_i = s_{i'}] = \frac{\sum_j \min\{\kappa_{ij}, \kappa_{i'j}\}}{\sum_j \max\{\kappa_{ij}, \kappa_{i'j}\}}.
\]

Our angular LSH module is outlined in Algorithm 3.3. Here we input an explicit “coarse” embedding \(f_i, c_i\) that we expect to be lower quality proxy of our target one. The coarse embedding can come from a weaker (and cheaper to train) model or from a partially-trained model. In our experiments we use a lower dimension SGNS model. Each LSH map is obtained by drawing a random vector and then mapping each entity \(i\) to the sign of a projection of \(f_i\) on the random vector. The probability that two focus entities have the same key depends on the angle between their coarse embedding vectors:

Lemma 3.3. \([3]\)

\[
\Pr[s_i = s_{i'}] = 1 - \frac{1}{\pi} \cos^{-1} \cos_{\text{sim}}(f_i, f_{i'}) = \frac{\|u \cdot v\|}{\|u\| \cdot \|v\|}
\]

where \(\cos_{\text{sim}}(v, u) := \frac{v \cdot u}{\|v\| \cdot \|u\|}\) is the cosine of the angle between the two vectors.

We can always apply multiple LSH maps to further refine basic microbatches. Each application decreases the microbatch size and increases quality (similarity level of entities placed in the same microbatch). More precisely, with
We select a row index $i$.

We perform experiments on synthetic data sets generated by the stochastic blocks model [6] to help us understand how a good embedding to have high cosine similarity for same-block pairs and low (around 0) cosine similarity was drawn. We performed experiments with several parameter choices and report representative results for squared $n$ groups of $B$.

 Eventually, we may hit a regime where IND arrangements dominate.

![Algorithm 2: Independent microbatch](image)

**Input:** $\kappa$  
Choose $(i, j)$ with probability $\kappa_{ij} / \| \kappa \|_1$:

```
foreach context j do
    M_j ← max_k $\kappa_{ij}$ // Maximum entry for context $j$
    Index column $j$ so that we can return for each $t \in (0, 1], P(j, t) := \{ i \mid \kappa_{ij} \geq tM_j \}$.
```

![Algorithm 3: Basic coordinated microbatches (Focus updates)](image)

**Input:** $\kappa$  
// Preprocessing:
```
foreach context $j$ do
    M_j ← max_k $\kappa_{ij}$ // Maximum entry for context $j$
    Index column $j$ so that we can return for each $t \in (0, 1], P(j, t) := \{ i \mid \kappa_{ij} \geq tM_j \}$.
```

// Microbatch draw:
```
Choose a context $j$ with probability $\frac{M_j}{\sum_h M_h}$
Draw $u \sim U[0, 1]$
return $P(j, u)$
```

![Algorithm 4: Jaccard LSH (Focus updates)](image)

**Input:** $\{ f_i \}$ // coarse $d$ dimensional embedding
```
foreach focus $i$ do // assign LSH bucket key
    $s_i \leftarrow \arg \min_j u_j / \kappa_{ij}$
```

![Algorithm 5: Angular LSH (Focus updates)](image)

**Input:** $\{ f_i \}$ // coarse $d$ dimensional embedding
```
foreach focus $i$ do // assign LSH bucket key
    $s_i \leftarrow \text{sign}(r \cdot f_i)$
```

4 Stochastic blocks data and quality measures

We perform experiments on synthetic data sets generated by the stochastic blocks model [6] to help us understand how the benefits of different arrangement methods vary and balance out by the stage of training and properties of the data. The parameters for the generative model are the dimensions $n \times n$ of the matrix, the number of (equal size) blocks $B$, the number of interactions $r$, and the in-block probability $p$. The rows and columns are partitioned to consecutive groups of $n/B$, where the $i$th part of rows and $i$th part of columns are considered to belong to the same block. We generate the matrix by initializing the associations to be $\kappa_{ij} = 0$. We then draw $r$ interactions independently as follows. We select a row index $i \in [n]$ uniformly at random. With probability $p$, we select (uniformly at random) a column $j \in [n]$ that is in the same block as $i$. Otherwise (with probability $1 - p$) we select a uniform column $j \in [n]$ that is outside the block of $i$. We then increment $\kappa_{ij}$. The final association $\kappa_{ij}$ is the number of times the interaction $(i, j)$ was drawn. We performed experiments with several parameter choices and report representative results for squared matrices with parameters $n = 10^4$, $r = 10^7$, $p = 0.7$ and $B \in \{ 10, 20, 50, 100 \}$. We used polynomially-decaying learning rate with a range of parameters using the TensorFlow library [1]. The learning rate parameters were tuned to perform well with independent arrangements. We worked with minibatch sizes of $b \in \{ 4, 64, 246 \}$ and trained embeddings of dimension $d \in \{ 3, 5, 10, 25, 50, 100 \}$.

We use two measures of the quality of an embedding with respect to the blocks ground truth. The first is the cosine gap which measures average quality and is defined as the difference in the average cosine similarity between positive examples and negative examples. We generate a sampled set $T_+$ of same-block pairs $(i, j)$ as positive test examples and a sampled set $T_-$ of pairs that are not in the same block as negative test examples and compute

$$
\frac{1}{|T_+|} \sum_{(i,j) \in T_+} \cos\text{Sim}(f_i, c_j) - \frac{1}{|T_-|} \sum_{(i,j) \in T_-} \cos\text{Sim}(f_i, c_j). \tag{1}
$$

We expect a good embedding to have high cosine similarity for same-block pairs and low (around 0) cosine similarity for out of block pairs. The second measure we use, precision at $k$, is focused on the quality of the top predictions and
We highlight two properties of coordinated arrangements that are beneficial to accelerating convergence: A micro-level property that makes gradient updates more effective by moving embedding vectors of similar entities closer and a macro-level property of preserving expected similarity in fractions of epochs.

**Effectiveness of gradient updates**  When updates on corresponding associations of two entities are processed in the same minibatch then the cosine similarity of their embedding vectors increases. This holds also in early training when the embedding vectors are randomly initialized. Similar entities have more corresponding associations (fraction equals the Jaccard similarity) and benefit more from this property. In particular, the SGNS loss term for a positive example is \( L_+(f, c) = \log \sigma(f, c) = \log \left(\frac{1}{1 + \exp(-f \cdot c)}\right) \). The gradient with respect to \( f \) is \( \nabla_f (L_+(f, c)) = c \cdot \frac{1}{1 + \exp(f \cdot c)} \) and the respective update of \( f' \leftarrow f + \eta \frac{1}{1 + \exp(f \cdot c)} c \) clearly increases \( \cos_{\text{sim}}(f, c) \). Consider two focus entities 1, 2 and corresponding positive associations with context entity \( j \). When positive examples \((1, j)\) and \((2, j)\) are in the same focus-updating minibatch, both \( \cos_{\text{sim}}(f_1, c) \) and \( \cos_{\text{sim}}(f_2, c) \) increase and a desirable side effect is that in expectation \( \cos_{\text{sim}}(f_1, f_2) \) increases as well. This is achieved when the updates on corresponding examples \((1, j)\) and \((2, j)\) is performed with the same current parameters \( c_j \), which happens with COO (and with full gradients) but less so with IND arrangements that on average place the two examples half an epoch apart. Figure 1 shows the expected increase in cosine similarity \( E[\cos_{\text{sim}}(f'_1, f'_2) - \cos_{\text{sim}}(f_1, f_2)] \) for learning rates \( \eta = 0.02, 0.05 \) when the vectors \( f_1, f_2, \) and \( c \) are independently drawn from a product distribution \( \mathcal{N}(0, 1)^d \) of independent Gaussians.

**Preserving similarity**  Coordinated arrangements preserve information on entity similarity in fraction of epochs. More formally, consider for two entities the weighted Jaccard similarity computed from examples in a small stretch of training that includes a very small number of examples with each entity. With COO, the expected similarity is equal to that of the full vectors whereas with IND, the similarity information disperses rapidly. This is because COO uses coordinated samples which maximize preserved similarity for the marginal distribution. It is instructive to consider two focus entities with Jaccard similarity \( J \) and \( M \) contexts with positive \( \kappa_{ij} = c > 0 \). An \( \alpha \ll 1 \) fraction of an epoch will on average include \( \alpha M \) sampled contexts from each focus entity. When the samples are independent then the sets would be highly dissimilar even when \( J \) is close to 1. When the samples are coordinated then the expected similarity in the sample corresponds to the similarity of the original vectors. We next demonstrate the similarity-preservation quality experimentally with stochastic block matrices. We select small sets of positive training examples using independent and coordinated sampling schemes according to the same per-entity marginal distributions. We then train with this
small set on multiple epochs until convergence as a way to gauge the “information” each set provides and its effect on training speed. We sample \( T = 5, 10, 15, 20 \) example interactions from each row (for focus updates) and symmetrically from each column (for context updates) of the association matrix. With independent sampling we select \( T \) independent examples for each row \( i \) by selecting a column \( j \) with probability \( \kappa_{ij}/\|\kappa_i\|_1 \). For coordinated sampling we repeat the following \( T \) times. We draw \( u_j \sim \text{Exp}[1] \) for each column and select for each \( i \) column \( \arg \max_j \kappa_{ij}/u_j \). Clearly the marginal distribution is the same, as the probability that column \( j \) is selected for row \( i \) is equal to \( \kappa_{ij}/\|\kappa_i\|_1 \). Symmetric schemes apply to columns. We trained embeddings (with IND arrangements) on these smaller sets of examples on otherwise identical setups. Training was one-sided and alternated on each minibatch with row samples used for updating row embeddings and column samples for updating column embeddings. Representative results \((b = 4)\) are reported in Figure 2. We observe that the coordinated selection of training examples consistently attains faster convergence in the earlier epochs. With fewer examples per entity, coordinated selection also had a higher peak quality than the respective independent selection. With more examples and larger blocks, the coordinated selection peaked lower, due to loss of the multi-hop expander structure.

6 Arrangement Methods Experiments

We train embeddings with different minibatch arrangement methods: The baseline independent arrangements (IND) as in Algorithm 2, coordinated arrangements with basic microbatches (COO) as in Algorithm 6, and coordinated arrangements with LSH partitioned microbatches (COO+LSH). Our experiments use COO+LSH arrangements with Jaccard and with angular applied with an embedding computed with low dimension \((d = 3)\). We use Jaccard COO+LSH with a single map (“Jaccard”) and we use both LSH functions with adaptive partition using independent maps until the microbatch size is below the minibatch size \( b \) (“Jaccard*” and “angular*”). We also evaluate adaptive arrangement (MIX) that start with COO, may switch to (one variant) of COO+LSH or to IND, and may switch from COO+LSH to IND. The switch points of MIX were determined once via a hyperparameter search and then used across repetitions (generated synthetic data and splits for recommendation data).

![Figure 3: Precision at \( k = 10 \) with different arrangement methods in the course of training \((d = 50, b = 64)\). Using \( 10^4 \times 10^4 \) stochastic blocks matrices with \( B \in \{10, 20, 50, 100\} \). The switch point for the MIX method are shown in blue (to COO+LSH) and green (to IND). The solid lines are for Jaccard LSH and the dashed lines are for angular LSH.](image)

6.1 Stochastic blocks

Representative results are reported for \( d = 50 \) and \( b = 64 \) and varying block sizes for the precision quality measure in Figure 3 and for the cosine gap quality measure in Figure 4. For each configuration, we show quality in the full course of training and also zoom on the early part of training.
The first thing to observe is that across all block sizes $B$ and for the two quality measures our coordinated arrangement methods result in faster convergence than the baseline IND methods. The zoom on early training reveals that COO is consistently the dominant arrangement method in the early regime but performance may deteriorate later in training – this is due to the shifting balance between the benefit of helpful co-placements and the cost of unhelpful co-placements that needlessly increase effective minibatch size. We observe that performance of COO on larger blocks ($B = 10$) deteriorates earlier than with small blocks ($B = 100$) – this is because matrices with larger blocks have larger microbatch sizes (and effective minibatch size). We can also observe that in the mid-training regime COO+LSH in particular the variants where the microbatch size never exceeds the minibatch size $b$ are dominant. This is because it retains many helpful co-placements without increasing effective minibatch size. The Jaccard COO+LSH that uses a single map does not perform as well with large blocks because microbatch sizes tend to be much larger than $b$. Finally we can see that our MIX arrangement methods yield improvements over respective pure methods. We report the results of additional experiments demonstrating consistent training gains of 5-30% over the baseline IND arrangements in Appendix A. These experiments also demonstrate that (as expected) the gain increases with minibatch size.

Figure 4: Cosine gap with different arrangement methods in the course of training ($d = 50, b = 64$). Using $10^4 \times 10^4$ stochastic blocks matrices with $B \in \{10, 20, 50, 100\}$. The switch point for the MIX method are shown in blue (to COO+LSH) and green (to IND). The solid lines are for Jaccard LSH and the dashed lines are for angular* LSH.

### 6.2 Recommendation data sets

| LSH | 0.75 × peak | 0.95 × peak | 0.99 × peak |
|-----|-------------|-------------|-------------|
| %gain | $\times 10^6$ | $\times 10^6$ | $\times 10^6$ |
| AMazon | Gain of COO + LSH over IND (peak = 3.4) |
| Jac | 4.29 | 3.50 | 6.86 | 5.83 | 11.02 | 7.17 |
| Ang | 10.00 | 3.50 | 13.38 | 5.83 | 16.04 | 7.17 |
| MOVIELENS1M: Gain of MIX over IND (peak = 0.40) |
| Jac | 2.13 | 1.41 | 0.58 | 1.73 | 1.55 | 1.93 |
| Ang | 4.96 | 1.41 | 8.67 | 1.73 | 11.92 | 1.93 |

Table 1: AMazon and MOVIELENS1M: Training gain over IND baseline ($b = 64$, cosine gap).

We performed experiments on two recommendation data sets, MOVIELENS1M and AMAZON. The MOVIELENS1M dataset [18] contains $10^6$ reviews by $6 \times 10^3$ users of $4 \times 10^3$ movies. The AMAZON dataset [25] contains $5 \times 10^5$ fine food reviews of $2.5 \times 10^5$ users on $7.5 \times 10^3$ food items. Provided review scores were [1-5] and we preprocessed the matrix by taking $\kappa_{ij}$ to be 1 for review score that is at least 3 and 0 otherwise. We then reweighed entries in the MOVIELENS1M dataset by dividing the value by the sum of its row and column to the power of 0.75. This is standard processing that retains only positive ratings and reweights to prevent domination of frequent entities.

We created a test set $T_+$ of positive examples by sampling 20% of the non zero entries with probabilities proportional to $\kappa_{ij}$. The remaining examples were used for training. As negative test examples $T_-$ we used random zero entries. We measured quality using the cosine gap equation [1] and show results averaged over 5 random splits of the data to
test and training sets and 5 runs per split. The MIX and COO+LSH Jaccard were the respective best performers on MOVIELENS1M and AMAZON. Training gains \(d = 50\) with respect to the IND baseline are reported in Table[1]. We observe consistent reduction in training which indicate that arrangement tuning is an effective tool also on these more complex real-life data sets.

7 Conclusion

We consider embedding computations with stochastic gradients and establish that the arrangement of training examples into minibatches can be a powerful performance knob. In particular, we introduced coordinated arrangements as a principled method to accelerate SGD training of embedding vectors. Our experiments focused on the popular SGNS loss and our methods were designed for pairwise associations. In future we hope to explore the use of coordinated arrangement with other loss objectives, deeper networks, and more complex association structures.

References

[1] M. Abadi and et al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] G. Alain, A. Lamb, C. Sankar, A. C. Courville, and Y. Bengio. Variance reduction in SGD by distributed importance sampling. CoRR, abs/1511.06481, 2015.

[3] Michael W. Berry, Susan T. Dumais, and Gavin W. O’Brien. Using linear algebra for intelligent information retrieval. SIAM Review, 37(4):573–595, 1995.

[4] E. Cohen, H. Kaplan, and S. Sen. Coordinated weighted sampling for estimating aggregates over multiple weight assignments. VLDB, 2, 2009. full version: http://arxiv.org/abs/0906.4560.

[5] G. Cohen. Auxiliary problem principle and decomposition of optimization problem. J Optim Theory Appl, 32, 1980.

[6] A. Condon and R. M. Karp. Algorithms for graph partitioning on the planted partition model. Random Struct. Algorithms, 18(2), 2001.

[7] I. Csiszar and G. Tusnády. Information geometry and alternating minimization procedures. Statistics & Decisions: Supplement Issues, 1, 1984.

[8] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic indexing. Journal of the American Society for Information Science, 41(6):391–407, September 1990.

[9] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. J. Mach. Learn. Res., 12, 2011.

[10] Susan T. Dumais. Latent semantic indexing (LSI): TREC-3 report. In D. K. Harman, editor, The Third Text Retrieval Conference (TREC-3), pages 219–230, Gaithersburg, MD, 1995. U. S. Dept. of Commerce, National Institute of Standards and Technology.

[11] T. Fu and Z. Zhang. Cpsg-mcmc: Clustering-based preprocessing method for stochastic gradient mcmc. In AISTATS, 2017.

[12] R. Gemulla, E. Nijkamp, P. J. Haas, and Y. Sismanis. Large-scale matrix factorization with distributed stochastic gradient descent. In ACM KDD 2011. ACM, 2011.

[13] M. X. Goemans and D. P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. J. ACM, 42(6), 1995.
[14] Y. Hu, Y. Koren, and C. Volinsky. Collaborative filtering for implicit feedback datasets. In ICDM, 2008.

[15] Y. Koren. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In KDD, 2008.

[16] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. Computer, 42, 2009.

[17] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In NIPS, 2013.

[18] Movielen 1M Dataset. http://grouplens.org/datasets/movielens/1m/.

[19] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In EMNLP, 2014.

[20] B. Perozzi, R. Al-Rfou, and S. Skiena. Deepwalk: Online learning of social representations. In KDD. ACM, 2014.

[21] H. Robbins and D. O. Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In J. S. Rustagi, editor, Optimizing Methods in Statistics. Academic Press, 1971.

[22] R. Salakhutdinov, A. Mnih, and G. Hinton. Restricted boltzmann machines for collaborative filtering. In ICML. ACM, 2007.

[23] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. Information Processing & Management, 24(5):513 – 523, 1988.

[24] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2015.

[25] Stanford network analysis project. http://snap.stanford.edu.

[26] C. Zhang, H. Kjellström, and S. Mandt. Stochastic learning on imbalanced data: Determinantal point processes for mini-batch diversification. In UAI 2017, 2017.
A Additional Stochastic blocks experiments

Results of additional experiments on $10^4 \times 10^4$ stochastic blocks are reported in Table 2 (Jaccard LSH MIX with cosine gap), Table 3 (Jaccard LSH MIX with precision at $k = 10$), Table 4 (angular* LSH MIX computed with $d = 3$ embedding with cosine gap), and Table 5 (angular* LSH MIX computed with $d = 3$ embedding with precision at $k = 10$). We report results for different minibatch sizes $B$. The tables list the peak performance (cosine gap or precision) of MIX and the amount of training used by IND to reach 0.75, 0.95, or 0.99 of that peak. We also show the reduction in training that is gained when MIX is used instead of IND. Overall, we can see that MIX consistently had gains of 5-30% in the amount of training required to reach certain quality. The gain is larger with smaller blocks and also with larger minibatches. The latter is because coordinated arrangement work with effective minibatch size that is determined by the microbatch sizes and may be larger than $b$ whereas IND always uses minibatches of specified size $b$.

| #blocks | mbatch size | $B$ | 0.75 %gain $\times 10^6$ | %gain $\times 10^6$ | 0.95 %gain $\times 10^6$ | %gain $\times 10^6$ | 0.99 %gain $\times 10^6$ | %gain $\times 10^6$ |
|---------|-------------|-----|--------------------------|-----------------|--------------------------|-----------------|--------------------------|-----------------|
| 10      | 4           | 1.09| 6.58 3.04                | 4.90 3.67       | 4.76 4.20                |                 |                          |                 |
| 10      | 64          | 1.09| 9.87 3.04                | 8.45 3.67       | 7.35 4.22                |                 |                          |                 |
| 10      | 256         | 1.08| 20.20 3.07               | 17.25 3.71      | 16.32 4.29               |                 |                          |                 |
| 20      | 64          | 1.03| 11.36 2.73               | 9.28 3.34       | 7.77 3.86                |                 |                          |                 |
| 20      | 256         | 1.03| 21.61 2.73               | 18.58 3.39      | 16.33 3.92               |                 |                          |                 |
| 50      | 4           | 1.00| 8.37 2.39                | 6.69 2.99       | 5.43 3.50                |                 |                          |                 |
| 50      | 64          | 1.00| 12.97 2.39               | 10.00 3.00      | 7.43 3.50                |                 |                          |                 |
| 50      | 256         | 1.00| 23.24 2.41               | 18.27 3.01      | 15.77 3.55               |                 |                          |                 |
| 100     | 4           | 0.99| 8.41 2.14                | 6.23 2.73       | 5.57 3.23                |                 |                          |                 |
| 100     | 64          | 0.99| 14.02 2.14               | 10.58 2.74      | 7.74 3.23                |                 |                          |                 |
| 100     | 256         | 0.99| 21.76 2.16               | 17.09 2.75      | 15.50 3.29               |                 |                          |                 |

Table 2: Training gain of Jaccard MIX arrangement with respect to IND arrangement baseline for $10^4 \times 10^4$ stochastic blocks. Peak is maximum cosine gap quality for MIX. We report the number of training examples for IND to reach 75%, 95%, and 99% of peak with respective percent reduction in training with MIX.

| #blocks | mbatch size | $B$ | 0.75 %gain $\times 10^6$ | %gain $\times 10^6$ | 0.95 %gain $\times 10^6$ | %gain $\times 10^6$ | 0.99 %gain $\times 10^6$ | %gain $\times 10^6$ |
|---------|-------------|-----|--------------------------|-----------------|--------------------------|-----------------|--------------------------|-----------------|
| 10      | 4           | 1.00| 9.63 2.18                | 8.61 2.44       | 6.61 2.57                |                 |                          |                 |
| 10      | 64          | 1.00| 14.22 2.18               | 13.93 2.44      | 12.69 2.60               |                 |                          |                 |
| 10      | 256         | 1.00| 28.77 2.19               | 26.12 2.45      | 25.57 2.62               |                 |                          |                 |
| 20      | 4           | 1.00| 15.50 2.00               | 15.25 2.23      | 13.14 2.36               |                 |                          |                 |
| 20      | 64          | 1.00| 29.15 1.99               | 26.46 2.23      | 24.79 2.38               |                 |                          |                 |
| 50      | 4           | 1.00| 9.50 1.79                | 8.04 1.99       | 5.77 2.08                |                 |                          |                 |
| 50      | 64          | 1.00| 15.64 1.79               | 14.07 1.99      | 10.58 2.08               |                 |                          |                 |
| 50      | 256         | 1.00| 28.89 1.80               | 26.37 2.01      | 23.08 2.08               |                 |                          |                 |
| 100     | 4           | 1.00| 10.30 1.65               | 7.65 1.83       | 3.66 1.91                |                 |                          |                 |
| 100     | 64          | 1.00| 18.18 1.65               | 14.21 1.83      | 11.40 1.93               |                 |                          |                 |
| 100     | 256         | 1.00| 28.31 1.66               | 24.32 1.85      | 21.32 1.97               |                 |                          |                 |

Table 3: Training gain of Jaccard MIX arrangement with respect to IND baseline for $10^4 \times 10^4$ stochastic blocks. Peak is maximum precision at $k = 10$ quality for MIX. We report the number of training examples for IND to reach 75%, 95%, and 99% of peak with respective percent reduction in training with MIX.

B Embedding dimension analysis

We report here results on the effect of the dimension on the embedding quality and convergence, focusing on training with IND arrangements. Figure 6 shows quality in the course of training for different dimensions for selected $10^4 \times 10^4$ stochastic block matrices. We show both the cosine gap and the precision with $k = 10$. Figure 5 shows the cosine gap and precision with $k = 10$ quality for the MOVIELENS1M and AMAZON data sets.

The precision on the recommendation data sets is computed over focus entities (users) with at least 20 positive entries. The precision is the fraction of top $k$ that are in the test set.
Table 4: Training gain of angular* LSH MIX arrangement (based on $d = 3$ embeddings) with respect to IND baseline for $10^4 \times 10^4$ stochastic blocks. Peak is maximum cosine gap quality for MIX. We report the number of training examples for IND to reach 75%, 95%, and 99% of peak with respective percent reduction in training with MIX.

| #blocks | mbatch | $B$ size | $b$ peak | 0.75 | 0.95 | 0.99 |
|---------|--------|----------|----------|------|------|------|
| 10      | 4      | 1.09     | 7.24     | 3.04 | 5.72 | 3.67 |
| 10      | 64     | 1.09     | 8.55     | 3.04 | 6.81 | 3.67 |
| 10      | 256    | 1.08     | 13.31    | 3.07 | 12.40| 3.71 |
| 20      | 4      | 1.03     | 8.42     | 2.73 | 5.99 | 3.34 |
| 20      | 64     | 1.03     | 9.16     | 2.73 | 7.19 | 3.34 |
| 20      | 256    | 1.03     | 17.95    | 2.73 | 15.04| 3.39 |
| 50      | 4      | 1.00     | 9.62     | 2.39 | 7.69 | 2.99 |
| 50      | 64     | 1.00     | 12.97    | 2.39 | 9.67 | 3.00 |
| 50      | 256    | 1.00     | 17.95    | 2.39 | 15.04| 3.39 |
| 100     | 4      | 0.99     | 10.28    | 2.14 | 8.06 | 2.73 |
| 100     | 64     | 0.99     | 14.49    | 2.14 | 10.95| 2.74 |
| 100     | 256    | 0.99     | 22.69    | 2.16 | 17.82| 2.75 |

Table 5: Training gain of angular* LSH MIX arrangement (based on $d = 3$ embeddings) with respect to IND baseline for $10^4 \times 10^4$ stochastic blocks. Peak is maximum precision at $k = 10$ quality for MIX. We report the number of training examples for IND to reach 75%, 95%, and 99% of peak with respective percent reduction in training with MIX.

| #blocks | mbatch | $B$ size | $b$ peak | 0.75 | 0.95 | 0.99 |
|---------|--------|----------|----------|------|------|------|
| 10      | 4      | 1.00     | 10.55    | 2.18 | 9.84 | 2.44 |
| 10      | 64     | 1.00     | 11.47    | 2.18 | 10.66| 2.44 |
| 10      | 256    | 1.00     | 23.29    | 2.19 | 22.45| 2.45 |
| 20      | 4      | 1.00     | 11.00    | 2.00 | 9.87 | 2.23 |
| 20      | 64     | 1.00     | 12.50    | 2.00 | 11.66| 2.23 |
| 20      | 256    | 1.00     | 24.62    | 1.99 | 23.32| 2.23 |
| 50      | 4      | 1.00     | 12.29    | 1.79 | 11.06| 1.99 |
| 50      | 64     | 1.00     | 16.20    | 1.79 | 14.57| 1.99 |
| 50      | 256    | 1.00     | 28.31    | 1.66 | 24.32| 1.85 |
| 100     | 4      | 1.00     | 12.12    | 1.65 | 9.84 | 1.83 |
| 100     | 64     | 1.00     | 18.18    | 1.65 | 15.30| 1.83 |
| 100     | 256    | 1.00     | 28.31    | 1.66 | 24.32| 1.85 |

Figure 5: Training (IND with $b = 64$) with different dimensions. From left: MOVIELENS1M (cosine gap and precision for $k = 50$) and AMAZON (cosine gap and precision for $k = 50$).
On all data sets we can observe slightly faster convergence with higher dimension in terms of number of training examples. The per-example training cost, however, increases much faster and proportionally to the dimension. This means that lower dimension is more effective in reaching a particular lower quality level. On the recommendations data sets and for the precision quality measure on the stochastic blocks data we can see that the peak quality increases with the dimension. In particular, we can see that the peak quality for $d = 3$ is considerably lower than for $d = 50$. This means that higher dimension are effective in providing better peak quality. This supports our use in the experiments of the $d = 3$ embedding at the basis of angular* COO+LSH microbatches in order to accelerate the training of $d = 50$ embeddings, which are costlier to train but provide higher peak quality.