

---

# Improving Optimization in Models With Continuous Symmetry Breaking

---

Robert Bamler<sup>1</sup> Stephan Mandt<sup>1</sup>

## Abstract

Many loss functions in representation learning are invariant under a continuous symmetry transformation. For example, the loss function of word embeddings (Mikolov et al., 2013b) remains unchanged if we simultaneously rotate all word and context embedding vectors. We show that representation learning models for time series possess an approximate continuous symmetry that leads to slow convergence of gradient descent. We propose a new optimization algorithm that speeds up convergence using ideas from gauge theory in physics. Our algorithm leads to orders of magnitude faster convergence and to more interpretable representations, as we show for dynamic extensions of matrix factorization and word embedding models. We further present an example application of our proposed algorithm that translates modern words into their historic equivalents.

## 1. Introduction

Symmetries frequently occur in machine learning. They express that the loss function of a model is invariant under a certain group of transformations. For example, the loss function of matrix factorization or word embedding models remains unchanged if we simultaneously rotate all embedding vectors with the same rotation matrix. This is an example of a *continuous* symmetry, since the rotations are parameterized by a continuum of real-valued angles.

Sometimes, the symmetry of a loss function is broken, e.g., due to the presence of an additional term that violates the symmetry. For example, this may be a weak regularizer. In this paper, we show that such symmetry breaking may induce slow convergence problems in gradient descent, in particular when the symmetry breaking is weak. We solve this problem with a new optimization algorithm.

---

<sup>1</sup>Los Angeles, CA, USA. Correspondence to: Robert Bamler <robert.bamler@gmail.com>, Stephan Mandt <stephan.mandt@gmail.com>.

Weak continuous symmetry breaking leads to an ill-conditioned optimization problem. When a loss function is invariant under a continuous symmetry, it has a manifold of *degenerate* (equivalent) minima. This is usually not a problem because any such minimum is a valid solution of the optimization problem. However, adding a small symmetry breaking term to the loss function lifts the degeneracy and forces the model to prefer one minimum over all others. As we show, this leads to an ill-conditioned Hessian of the loss, with a small curvature along symmetry directions and a large curvature perpendicular to them. The ill-conditioned Hessian results in slow convergence of gradient descent.

We propose an optimization algorithm that speeds up convergence by separating the optimization in the symmetry directions from the optimization in the remaining directions. At regular intervals, the algorithm efficiently minimizes the small symmetry breaking term in such a way that the minimization does not degrade the symmetry invariant term.

Symmetries can be broken explicitly, e.g., due to an additional term such as an  $L_1$  regularizer in a word embedding model (Sun et al., 2016). However, perhaps more interestingly, symmetries can also be broken by couplings between model parameters. This is known as *spontaneous* symmetry breaking in the physics community.

One of our main findings is that spontaneous symmetry breaking occurs in certain time series models, such as dynamic matrix factorizations and dynamic word embedding models (Lu et al., 2009; Koren, 2010; Charlin et al., 2015; Bamler & Mandt, 2017; Rudolph & Blei, 2017). In these models, it turns out that model parameters may be twisted along the time axis, and that these twists contribute only little to the loss, thus leading to a small gradient. These low cost configurations are known in the physics community as Goldstone modes (Altland & Simons, 2010).

Our contributions are as follows:

- We identify a broad class of models that suffer from slow convergence of gradient descent due to Goldstone modes. We explain both mathematically and pictorially how Goldstone modes lead to slow convergence.
- Using ideas from gauge theories in physics, we propose Goldstone Gradient Descent (Goldstone-GD), an

optimization algorithm that speeds up convergence by separating the optimization along symmetry directions from the remaining coordinate directions.

- We evaluate the Goldstone-GD algorithm experimentally with dynamic matrix factorizations and Dynamic Word Embeddings. We find that Goldstone-GD converges orders of magnitude faster and finds more interpretable embedding vectors than standard gradient descent (GD) or GD with diagonal preconditioning.
- For Dynamic Word Embeddings (Bamler & Mandt, 2017), Goldstone-GD allows us to find historic synonyms of modern English words, such as “wagon” for “car”. Without our advanced optimization algorithm, we were not able to perform this task.

Our paper is structured as follows. Section 2 describes related work. In Section 3, we specify the model class under consideration, introduce concrete example models, and discuss the slow convergence problem. In Section 4, we propose the Goldstone-GD algorithm that solves the slow convergence problem. We report experimental results in Section 5 and provide concluding remarks in Section 6.

## 2. Related Work

Our paper discusses continuous symmetries in machine learning and proposes a new optimization algorithm. In this section, we summarize related work on both aspects.

Most work on symmetries in machine learning focuses on discrete symmetries. Convolutional neural networks (LeCun et al., 1998) exploit the discrete translational symmetry of images. This idea was generalized to arbitrary discrete symmetries (Gens & Domingos, 2014), to the permutation symmetry of sets (Zaheer et al., 2017), and to discrete symmetries in graphical models (Bui et al., 2012; Noessner et al., 2013). Discrete symmetries do not cause an ill-conditioned optimization problem because they lead to isolated degenerate minima rather than a manifold of degenerate minima.

In this work, we consider models with continuous symmetries. A specialized optimization algorithm for a loss function with a continuous symmetry was presented in (Choi et al., 1999). Our algorithm applies to a broader class of models since our assumptions are more permissive. We only require invariance under a collective rotation of all feature vectors, and not under independent symmetry transformations of each individual feature. Rotational symmetries have been identified in deep neural networks (Badrinarayanan et al., 2015), matrix factorization (Mnih & Salakhutdinov, 2008; Gopalan et al., 2015), linear factor models (Murphy, 2012), and word embeddings (Mikolov et al., 2013a;b; Pennington et al., 2014; Barkan, 2017). Dynamic matrix factorizations (Lu et al., 2009; Koren, 2010; Sun et al., 2012;

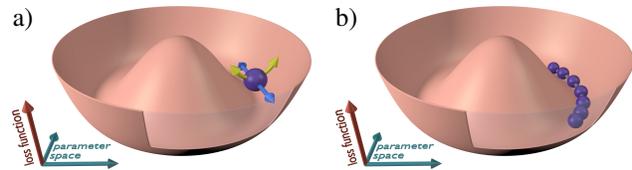


Figure 1. a) The loss  $\ell$  of a rotationally symmetric model has a continuum of degenerate minima. b) A Goldstone mode; similar configurations arise during the optimization of time series models, and they lead to a slowdown of gradient descent (see Figure 2).

Charlin et al., 2015) and dynamic word embeddings (Bamler & Mandt, 2017; Rudolph & Blei, 2017) generalize these models to sequential data. These are the models whose optimization we address in this paper.

The slow convergence in these models is caused by shallow directions of the loss function. Popular methods to escape a shallow valley of a loss function (Duchi et al., 2011; Zeiler, 2012; Kingma & Ba, 2014) use diagonal preconditioning. As confirmed by our experiments, diagonal preconditioning does not speed up convergence when the shallow directions correspond to collective rotations of many model parameters, which are not aligned with the coordinate axes.

Natural gradients (Amari, 1998; Martens, 2014) are a more sophisticated form of preconditioning, which has been applied to deep learning (Pascanu & Bengio, 2013) and to variational inference (Hoffman et al., 2013). Our proposed algorithm uses natural gradients in a subspace where they are cheap to obtain. Different to the Krylov subspace method (Vinyals & Povey, 2012), we construct the subspace such that it always contains the shallow directions of the loss.

## 3. Problem Setting

In this section, we formalize the notion of continuous symmetry breaking (Section 3.1), and we specify the type of models that we investigate in this paper (Section 3.2). We then show that the introduced models exhibit a specific type of symmetry breaking that is generically weak and therefore leads to slow convergence of gradient descent (Section 3.3).

### 3.1. Symmetry Breaking in Representation Learning

We formalize the concept of weak continuous symmetry breaking and show that it leads to an ill-conditioned optimization problem. Figure 1a illustrates the main effects of a continuous symmetry by means of a small toy model. The red surface depicts the loss function. Since the loss is rotationally symmetric, it has a ring of degenerate (i.e., equivalent) minima. The purple sphere depicts one arbitrarily chosen minimum. Notice that the loss is flat in the direction tangential to the ring of degenerate minima (blue arrows), i.e., its curvature along this direction is zero.

For a machine learning example, consider factorizing a large matrix  $X$  into the product  $U^\top V$  of two smaller matrices  $U$  and  $V$  by minimizing the loss  $\ell(U, V) = \|X - U^\top V\|_2^2$ . Rotating all columns of  $U$  and  $V$  by the same orthogonal<sup>1</sup> rotation matrix  $R$  such that  $U \leftarrow RU$  and  $V \leftarrow RV$  does not change  $\ell$  since  $(RU)^\top RV = U^\top (R^\top R)V = U^\top V$ .

The continuous rotational symmetry of  $\ell$  leads to a manifold of degenerate minima, and to a zero eigenvalue of the Hessian of  $\ell$  at any of the minima. If  $(U^*, V^*)$  minimizes  $\ell$ , then so does  $(RU^*, RV^*)$  for any rotation matrix  $R$ . On the manifold of degenerate minima, the gradient assumes the constant value of zero. A constant gradient (first derivative) means that the curvature (second derivative) is zero. More precisely, the Hessian of  $\ell$  has a zero eigenvalue for all eigenvectors that are tangential to the manifold of degenerate minima. Usually, a zero eigenvalue of the Hessian indicates a maximally ill-conditioned optimization problem, but this is not an issue here. The zero eigenvalue only means that convergence within the manifold of degenerate minima is infinitely slow. This is of no concern since any minimum is a valid solution of the optimization problem.

A problem arises when the continuous symmetry is weakly broken, e.g., by adding a small  $L_1$  regularizer to the loss. A sufficiently small regularizer changes the eigenvalues of the Hessian only slightly, leaving it still poorly conditioned. However, even if the regularizer has a tiny coefficient, it lifts the degeneracy and turns the manifold of exactly degenerate minima into a shallow valley with one preferred minimum. Convergence along this shallow valley is slow because of the small eigenvalues of the Hessian. This is the slow convergence problem that we address in this paper. We present a more natural setup in which the problem occurs in Sections 3.2-3.3. Our solution, presented in Section 4, is to separate the optimization in the symmetry directions from the optimization in the remaining directions.

### 3.2. Representation Learning for Time Series

We define a broad class of representation learning models for sequential data, and introduce three example models that are investigated experimentally in this paper. As we show in Section 3.3, the models presented here suffer from slow convergence due to a specific kind of symmetry breaking.

**General Model Class.** We consider data  $\mathbf{X} \equiv \{X_t\}_{t=1:T}$  that are associated with additional metadata  $t$ , such as a time stamp. For each  $t$ , the task is to learn a low dimensional representation  $Z_t$  by minimizing what we call a ‘local loss function’  $\ell(X_t; Z_t)$ . We add a quadratic regularizer  $\psi(\mathbf{Z})$  that couples the representations  $\mathbf{Z} \equiv \{Z_t\}_{t=1:T}$  along the  $t$ -dimension. In a Bayesian setup,  $\psi$  comes from the log-

prior of the model. The total loss function is thus

$$\mathcal{L}(\mathbf{Z}) = \sum_{t=1}^T \ell(X_t; Z_t) + \psi(\mathbf{Z}). \quad (1)$$

For each task  $t$ , the representation  $Z_t$  is a matrix whose columns are low dimensional embedding vectors. We assume that  $\ell$  is invariant under a collective rotation of all columns of  $Z_t$ : let  $R$  be an arbitrary orthogonal rotation matrix of the same dimension as the embedding space, then

$$\ell(X_t; RZ_t) = \ell(X_t; Z_t). \quad (2)$$

Finally, we consider a specific form of the regularizer  $\psi$  which is quadratic in  $\mathbf{Z}$ , and which is defined in terms of a sparse symmetric coupling matrix  $\mathbf{L} \in \mathbb{R}^{T \times T}$ :

$$\psi(\mathbf{Z}) = \frac{1}{2} \text{Tr}(\mathbf{Z}^\top \mathbf{L} \mathbf{Z}). \quad (3)$$

Here, the matrix-vector multiplications are carried out in  $t$ -space, and the trace runs over the remaining dimensions. Note that, different to Section 3.1, we do not require  $\psi$  to have a small coefficient. We only require the coupling matrix  $\mathbf{L}$  to be sparse. In the examples below,  $\mathbf{L}$  is tridiagonal and results from a Gaussian Markovian time series prior. In a more general setup,  $\mathbf{L} = \mathbf{D} - \mathbf{A}$  is the Laplacian matrix of a sparse weighted graph (Poignard et al., 2018). Here,  $\mathbf{A}$  is the adjacency matrix, whose entries are the coupling strengths, and the degree matrix  $\mathbf{D}$  is diagonal and defined such that the entries of each row of  $\mathbf{L}$  sum up to zero.

Equations 1, 2, and 3 specify the problem class of interest in this paper. The following paragraphs introduce the specific example models used in our experiments. In Section 3.3, we show that the sparse coupling in these models leads to weak continuous symmetry breaking and therefore to slow convergence of gradient descent (GD).

**Model 1: Dense Dynamic Matrix Factorization.** Consider the task of factorizing a large matrix  $X_t$  into a product  $U_t^\top V_t$  of two smaller matrices. The latent representation is the concatenation of the two embedding matrices,

$$Z_t \equiv (U_t, V_t). \quad (4)$$

In a Gaussian matrix factorization, the local loss function is

$$\ell(X_t; Z_t) = -\log \mathcal{N}(X_t; U_t^\top V_t, I) \quad (5)$$

In dynamic matrix factorization models, the data  $\mathbf{X}$  are observed sequentially at discrete time steps  $t$ , and the representations  $\mathbf{Z}$  capture the temporal evolution of latent embedding vectors. We use a Markovian Gaussian time series prior with a coupling strength  $\lambda$ , resulting in the regularizer

$$\psi(\mathbf{Z}) = \frac{\lambda}{2} \sum_{i=1}^N \sum_{t=1}^{T-1} \|z_{t+1,i} - z_{t,i}\|_2^2. \quad (6)$$

<sup>1</sup>We call a square matrix  $R$  ‘orthogonal’ if  $R^\top R$  is the identity. This is sometimes also called an ‘orthonormal’ matrix.

Here, the vector  $z_{t,i}$  is the  $i^{\text{th}}$  column of the matrix  $Z_t$ , i.e., the  $i^{\text{th}}$  embedding vector, and  $N$  is the number of columns. The regularizer allows the model to share statistical strength across time. By multiplying out the square, we find that  $\psi$  has the form of Eq. 3 with a tridiagonal coupling matrix,

$$\mathbf{L} = \lambda \begin{pmatrix} 1 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \end{pmatrix}. \quad (7)$$

**Model 2: Sparse Dynamic Matrix Factorization.** In a sparse matrix factorization, the local loss  $\ell$  involves only few components of the matrix  $U_t^\top V_t$ . The latent representation is again  $Z_t \equiv (U_t, V_t)$ . We consider a model for movie ratings where each user rates only few movies. When user  $i$  rates movie  $j$  in time step  $t$ , we model the log-likelihood to obtain the binary rating  $x \in \{\pm 1\}$  with a logistic regression,

$$\log p(x|u_{t,i}, v_{t,j}) = \log \sigma(x u_{t,i}^\top v_{t,j}) \quad (8)$$

with the sigmoid function  $\sigma(\xi) = 1/(1 + e^{-\xi})$ . Eq. 8 is the log-likelihood of the rating  $x$  of a single movie by a single user. We obtain the full log-likelihood  $\log p(X_t|Z_t)$  for time step  $t$  by summing over the log-likelihoods of all ratings observed at time step  $t$ . The local loss is

$$\ell(X_t; Z_t) = -\log p(X_t|Z_t) + \frac{\gamma}{2} \|Z_t\|_2^2. \quad (9)$$

Here,  $\|\cdot\|_2$  is the Frobenius norm, and we add a quadratic regularizer with strength  $\gamma$  since data for some users or movies may be scarce. We distinguish this local regularizer from the time series regularizer  $\psi$ , given again in Eq. 6, as the local regularizer does not break the rotational symmetry.

**Model 3: Dynamic Word Embeddings.** Word embeddings map words from a large vocabulary to a low dimensional representation space such that neighboring words are semantically similar, and differences between word embedding vectors capture syntactic and semantic relations. We consider the Dynamic Word Embeddings model (Bamler & Mandt, 2017), which uses a probabilistic interpretation of the Skip-Gram model with negative sampling, also known as word2vec (Mikolov et al., 2013b; Barkan, 2017), and combines it with a time series prior. The model is trained on  $T$  text sources with time stamps  $t$ , and it assigns two time dependent embedding vectors  $u_{t,i}$  and  $v_{t,i}$  to each word  $i$  from a fixed vocabulary. The embedding vectors are obtained by simultaneously factorizing two matrices, which contain so-called positive and negative counts of word-context pairs. Therefore, the representation  $Z_t \equiv (U_t, V_t)$  for each time step is invariant under orthogonal transformations. The regularizer  $\psi$  comes from a time series prior that is a discretized Ornstein-Uhlenbeck process, i.e., it combines a random diffusion process with a local quadratic regularizer.

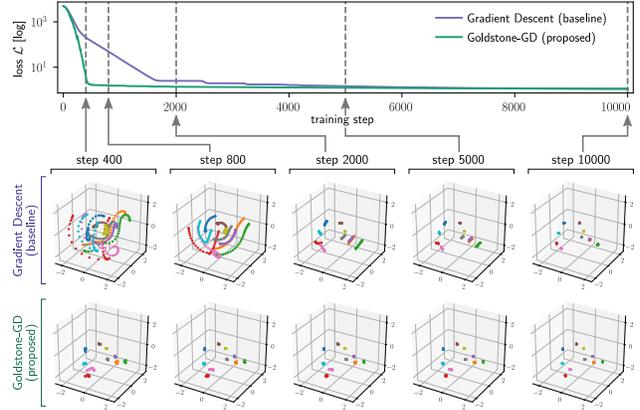


Figure 2. Goldstone modes and slow convergence of GD in a dynamic matrix factorization with a  $3d$  embedding space (details in Section 5.1). Colored points in the  $3d$  plots show the evolution of each embedding vector along the time dimension of the model.

### 3.3. Symmetry Breaking in Representation Learning for Time Series

We show that the time series models introduced in Section 3.2 exhibit a variant of the symmetry breaking discussed in Section 3.1. The symmetry breaking in time series models is generically weak, thus causing slow convergence.

**Geometric Picture: Goldstone Modes.** Figure 1b graphically illustrates the class of time series representation learning problems introduced in Eq. 1. Each of the seven purple spheres depicts an embedding vector  $Z_t$  with a time stamp  $t$  in a rotationally symmetric local loss function  $\ell$  (red surface). For a simpler visualization, we assume here that the local loss function is the same for every sphere, and that each  $Z_t$  contains only a single embedding vector.

The embedding vectors are coupled via the term  $\psi$  in Eq. 1. We can think of this coupling as springs between neighboring spheres (not drawn) that try to pull the spheres together. Figure 1b shows a typical configuration that GD may find after a few update steps. Each embedding  $Z_t$  minimizes its local loss  $\ell$ . However, the chain is not yet contracted to a single point, which reflects a deviation from the minimum of the total loss  $\mathcal{L}$ . Such a configuration is called a Goldstone mode in the physics literature (Altland & Simons, 2010).

As we show below, the gradient of the total loss is suppressed in a Goldstone mode (Goldstone theorem) (Altland & Simons, 2010). This leads to slow convergence of GD, as we illustrate in Figure 2. The  $3d$  plots show snapshots of the embedding space in a Gaussian dynamic matrix factorization with  $T = 30$  time stamps and a  $3d$  representation space (details in Section 5.1). Each color corresponds to an embedding vector. Points of equal color show the same embedding vector at different time stamps  $t$ .

In this toy experiment, the local loss  $\ell$  is again identical for all time stamps. Thus, in the optimum, the embedding vectors at different times should be the same, so that the trajectories contract to single points. We see that GD (top row of  $3d$  plots) takes a long time for the chains to contract, while our algorithm finds the optimum much faster.

**Formal Picture: Eigenvalues of the Hessian.** Goldstone modes decay slowly in GD because the gradient of the total loss  $\mathcal{L}$  is suppressed in a Goldstone mode. This can be seen by analyzing the eigenvalues of the Hessian  $\mathbf{H}$  of  $\mathcal{L}$  at its minimum. For a configuration  $\mathbf{Z}$  that is close to the true minimum  $\mathbf{Z}^*$ , the gradient is approximately  $\mathbf{H}(\mathbf{Z} - \mathbf{Z}^*)$ .

The Hessian of  $\mathcal{L}$  in Eq. 1 is the sum of the Hessians of the local loss functions  $\ell$  plus the Hessian of the regularizer  $\psi$ . As discussed in Section 3.1, the Hessians of  $\ell$  all have exact zero eigenvalues along the symmetry directions. Within this nullspace, only the Hessian  $\mathbf{H}^{(\psi)}$  of the regularizer  $\psi$  remains. From Eq. 3, we find  $\mathbf{H}^{(\psi)} = I_{N \times N} \otimes I_{d \times d} \otimes \mathbf{L}$  where  $\otimes$  is the tensor product, and  $I_{N \times N}$  and  $I_{d \times d}$  are identity matrices in the input and embedding space, respectively. Thus,  $\mathbf{H}^{(\psi)}$  has the same eigenvalues as the Laplacian matrix  $\mathbf{L}$  of the coupling graph, each with multiplicity  $Nd$ .

Since the rows of  $\mathbf{L}$  sum up to zero,  $\mathbf{L}$  has a zero eigenvalue for the eigenvector  $(1, \dots, 1)^\top$ . This is a consequence of the fact that, despite the symmetry breaking regularizer  $\psi$ , the total loss  $\mathcal{L}$  is still invariant if we rotate all embeddings  $\mathbf{Z}$  by the same orthogonal matrix. As discussed in Section 3.1, this remaining symmetry leads to zero eigenvalues of the Hessian, but these do not induce slow convergence of GD.

The speed of convergence of GD is governed by the lowest nonzero eigenvalue of the Hessian, and therefore of  $\mathbf{L}$ . In a Markovian time series model,  $\mathbf{L}$  in Eq. 7 couples only neighbors along a chain of length  $T$ . Its lowest nonzero eigenvalue is  $2\lambda(1 - \cos(\pi/T))$  (de Abreu, 2007), which vanishes as  $O(1/T^2)$  for large  $T$ . Thus, even if the coupling strength  $\lambda$  is large, the gradient of  $\psi$  is suppressed for the corresponding eigenvectors, i.e., the Goldstone modes. A more general model may couple tasks  $t$  along a sparse graph other than a chain. The second lowest eigenvalue of the Laplacian matrix  $\mathbf{L}$  of a graph is called the algebraic connectivity (de Abreu, 2007), and it is small in sparse graphs. This makes the optimization problem ill-conditioned.

## 4. Goldstone Gradient Descent

In this section, we present our solution to the slow convergence problem that we identified in Section 3.3. Algorithm 1 summarizes the proposed Goldstone Gradient Descent (Goldstone-GD) algorithm. We lay out details in Section 4.1, and discuss hyperparameters in Section 4.2.

The algorithm minimizes a loss function  $\mathcal{L}$  of the form of

---

### Algorithm 1: Goldstone Gradient Descent (Goldstone-GD)

---

**Input:** Loss function  $\mathcal{L}$  of the form of Eqs. 1-3; learning rate  $\rho$ ; integer hyperparameters  $k_1$  and  $k_2$

**Output:** Local minimum of  $\mathcal{L}$ .

```

1 Initialize model parameters  $\mathbf{Z}$  randomly
2 Initialize gauge fields  $\tilde{\Gamma} \leftarrow \mathbf{0}$ 
3 repeat
4   repeat  $k_1$  times
5     Set  $\mathbf{Z} \leftarrow \mathbf{Z} - \rho \nabla_{\mathbf{Z}} \mathcal{L}(\mathbf{Z})$ 
6      $\triangleright$  gradient step in full parameter space
7   end
8   Obtain  $\mathbf{M}$  and  $\rho'$  from Eqs. 15 and 17
9    $\triangleright$  transformation to symmetry subspace
10  repeat  $k_2$  times
11    Set  $\tilde{\Gamma} \leftarrow \tilde{\Gamma} - \rho' \mathbf{L}^+ \nabla_{\tilde{\Gamma}} \mathcal{L}'(\tilde{\Gamma}; \mathbf{M})$ 
12     $\triangleright$  natural gradient step in symmetry subspace
13  end
14  Set  $Z_t \leftarrow Z_t + (\tilde{\Gamma}_t - \tilde{\Gamma}_t^\top) Z_t \quad \forall t \in \{1, \dots, T\}$ 
15   $\triangleright$  transformation back to full parameter space
until convergence
    
```

---

Eqs. 1-3. The main idea is to periodically minimize the symmetry breaking regularizer  $\psi$  more efficiently than GD, without degrading the symmetry invariant part of the loss. We alternate between this specialized minimization of  $\psi$  (lines 7-8 in Algorithm 1) and standard GD on the full loss function  $\mathcal{L}$  (lines 4-5). Switching between the two types of updates involves an overhead due to coordinate transformations (lines 6 and 9). We therefore always perform several updates of each type in a row (hyperparameters  $k_1$  and  $k_2$ ). Algorithm 1 presents Goldstone-GD in its simplest form. It is straight-forward to combine it with adaptive learning rates and minibatch sampling, see experiments in Section 5.

### 4.1. Optimization in the Symmetry Subspace

We explain lines 6-9 of Algorithm 1. These steps minimize the total loss function  $\mathcal{L}(\mathbf{Z})$  while restricting updates of the model parameters  $\mathbf{Z}$  to symmetry transformations. Let  $\mathbf{R} \equiv \{R_t\}_{t=1:T}$  denote  $T$  orthogonal matrices. The task is to minimize the following auxiliary loss function over  $\mathbf{R}$ ,

$$\mathcal{L}'(\mathbf{Z}; \mathbf{R}) \equiv \mathcal{L}(R_1 Z_1, \dots, R_T Z_T) - \mathcal{L}(Z_1, \dots, Z_T) \quad (10)$$

with the nonlinear constraint  $R_t^\top R_t = I \forall t$ . If  $\mathbf{R}^*$  minimizes  $\mathcal{L}'$ , then updating  $Z_t \leftarrow R_t^* Z_t$  decreases the loss  $\mathcal{L}$  by eliminating all Goldstone modes. The second term on the right-hand side of Eq. 10 does not influence the minimization as it is independent of  $\mathbf{R}$ . Subtracting this term makes  $\mathcal{L}'$  independent of the local loss functions  $\ell$ : by using Eqs. 1-2, we can write  $\mathcal{L}'$  in terms of only the regularizer  $\psi$ ,

$$\mathcal{L}'(\mathbf{Z}; \mathbf{R}) = \psi(R_1 Z_1, \dots, R_T Z_T) - \psi(Z_1, \dots, Z_T). \quad (11)$$

**Artificial Gauge Fields.** We turn the constrained minimization of  $\mathcal{L}'$  over  $\mathbf{R}$  into an unconstrained minimization using a result from the theory of Lie groups (Hall, 2015). Every special orthogonal matrix  $R_t \in SO(d)$  is the matrix exponential of a skew symmetric  $d \times d$  matrix  $\Gamma_t$ . Here, skew symmetry means that  $\Gamma_t^\top = -\Gamma_t$ , and the matrix exponential function  $\exp(\cdot)$  is defined by its series expansion,

$$R_t = \exp(\Gamma_t) \equiv I + \Gamma_t + \frac{1}{2!}\Gamma_t^2 + \frac{1}{3!}\Gamma_t^3 + \dots \quad (12)$$

which is not to be confused with the componentwise exponential of  $\Gamma_t$  (the term  $\Gamma_t^2$  in Eq. 12 is the matrix product of  $\Gamma_t$  with itself, not the componentwise square). Eq. 12 follows from the Lie group–Lie algebra correspondence for the Lie group  $SO(d)$  (Hall, 2015). Note that  $R_t$  is close to the identity  $I$  if the entries of  $\Gamma_t$  are small. To enforce skew symmetry of  $\Gamma_t$ , we parameterize it via the skew symmetric part of an unconstrained  $d \times d$  matrix  $\tilde{\Gamma}_t$ , i.e.,

$$\Gamma_t = \tilde{\Gamma}_t - \tilde{\Gamma}_t^\top. \quad (13)$$

We call the components of  $\tilde{\Gamma} \equiv \{\tilde{\Gamma}_t\}_{t=1:T}$  the gauge fields, invoking an analogy to gauge theory in physics.

**Taylor Expansion in the Gauge Fields.** Eqs. 12-13 parameterize a valid rotation matrix  $R_t$  in terms of an arbitrary  $d \times d$  matrix  $\tilde{\Gamma}_t$ . This turns the constrained minimization of  $\mathcal{L}'$  into an unconstrained one. However, the matrix-exponential function in Eq. 12 is numerically expensive, and its derivative is complicated because the group  $SO(d)$  is non-abelian. We simplify the problem by introducing an approximate. As the model parameters  $\mathbf{Z}$  approach the minimum of  $\mathcal{L}$ , the optimal rotations  $\mathbf{R}^*$  that minimize  $\mathcal{L}'$  converge to the identity, and thus the gauge fields converge to zero. In this limit, the approximation becomes exact.

We approximate the auxiliary loss function  $\mathcal{L}'$  by a second order Taylor expansion  $\mathcal{L}''$ . In detail, we truncate Eq. 12 after the term quadratic in  $\Gamma_t$  and insert the truncated series into Eq. 11. We multiply out the quadratic form in the prior  $\psi$ , Eq. 3, and neglect again all terms of higher than quadratic order in  $\Gamma$ . Using the skew symmetry of  $\Gamma_t$  and the symmetry of the Laplacian matrix  $\mathbf{L} = \mathbf{D} - \mathbf{A}$ , we find

$$\mathcal{L}''(\tilde{\Gamma}; \mathbf{M}) = \sum_{t,t'} A_{tt'} \text{Tr} \left[ \left( \Gamma_{t'} + \frac{1}{2}(\Gamma_{t'} - \Gamma_t)\Gamma_t \right) M_{tt'} \right] \quad (14)$$

where the trace runs over the embedding space, and for each  $t, t' \in \{1, \dots, T\}$ , we define the matrix  $M_{tt'} \in \mathbb{R}^{d \times d}$ ,

$$M_{tt'} \equiv \sum_{i=1}^N z_{t,i} z_{t',i}^\top. \quad (15)$$

We evaluate the matrices  $M_{tt'}$  on line 6 in Algorithm 1. Note that the adjacency matrix  $\mathbf{A}$  is sparse, and that we only need to obtain those matrices  $M_{tt'}$  for which  $A_{tt'} \neq 0$ .

We describe the numerical minimization of  $\mathcal{L}''$  below. Once we obtain gauge fields  $\tilde{\Gamma}^*$  that minimize  $\mathcal{L}''$ , the optimal update step for the model parameters would be  $Z_t \leftarrow \exp(\tilde{\Gamma}_t^* - \tilde{\Gamma}_t^{*\top}) Z_t$ . For efficiency, we truncate the matrix exponential function  $\exp(\cdot)$  after the linear term, resulting in line 9 of Algorithm 1. We do not reset the gauge fields  $\tilde{\Gamma}$  to zero after updating  $\mathbf{Z}$ , so that the next minimization of  $\mathcal{L}''$  starts with preinitialized  $\tilde{\Gamma}$ . This turned out to speed up convergence in our experiments, possibly because  $\tilde{\Gamma}$  acts like a momentum in the symmetry subspace.

**Natural Gradients.** Lines 7-8 in Algorithm 1 minimize  $\mathcal{L}''$  over the gauge fields  $\tilde{\Gamma}$  using GD. We speed up convergence using the fact that  $\mathcal{L}''$  depends only on the prior  $\psi$  and not on  $\ell$ . Since we know the Hessian of  $\psi$ , we can use natural gradients (Amari, 1998), resulting in the update step

$$\tilde{\Gamma} \leftarrow \tilde{\Gamma} - \rho' \mathbf{L}^+ \nabla_{\tilde{\Gamma}} \mathcal{L}''(\tilde{\Gamma}; \mathbf{M}) \quad (16)$$

where  $\rho'$  is a constant learning rate and  $\mathbf{L}^+$  is the pseudoinverse of the Laplacian matrix  $\mathbf{L}$ . We obtain  $\mathbf{L}^+$  by taking the eigendecomposition of  $\mathbf{L}$  and inverting the eigenvalues, except for the single zero eigenvalue corresponding to (irrelevant) global rotations, which we leave at zero.  $\mathbf{L}^+$  has to be obtained only once before entering the training loop.

**Learning Rate.** We find that we can automatically set  $\rho'$  in Eq. 16 to a value that leads to fast convergence,

$$\rho' = \frac{1}{TN \langle Z^2 \rangle} \quad \text{with} \quad \langle Z^2 \rangle \equiv \frac{1}{TNd} \sum_{t,i} \|z_{t,i}\|_2^2. \quad (17)$$

We arrive at this choice of learning rate by estimating the Hessian of  $\mathcal{L}''$ . The preconditioning with  $\mathbf{L}^+$  in Eq. 16 takes into account the structure of the Hessian in  $t$ -space, which enters  $\mathcal{L}''$  in Eq. 14 via the adjacency matrix  $\mathbf{A}$ . The remaining factor  $M_{tt'}$ , defined in Eq. 15, is quadratic in the components of  $\mathbf{Z}$  and linear in  $N$ . This suggests a learning rate  $\rho' \propto 1/(N \langle Z^2 \rangle)$ . We find empirically for large models that the  $t$ -dependency of  $M_{tt'}$  leads to a small mismatch between  $\mathbf{L}$  and the Hessian of  $\mathcal{L}''$ . The more conservative choice of learning rate in Eq. 17 leads to fast convergence of the gauge fields in all our experiments.

## 4.2. Hyperparameters

Goldstone-GD has two integer hyperparameters,  $k_1$  and  $k_2$ , which control the frequency of execution of each operation. Table 1 lists the computational complexity of each operation, assuming that the sparse adjacency matrix  $\mathbf{A}$  has  $O(T)$  nonzero entries, as is the case in Markovian time series models (Eq. 7). Note that the embedding dimension  $d$  is typically orders of magnitude smaller than the input dimension  $N$ . Therefore, update steps in the symmetry subspace (line 8) are cheap. In our experiments, we always set  $k_1$  and  $k_2$  such that the overhead from lines 6-9 is less than 10%.

Table 1. Runtimes of operations in Goldstone-GD (L=line in Algorithm 1; #=frequency of execution;  $T$ =no. of time steps;  $N$ =input dimension;  $d$ =embedding dimension;  $k_1, k_2$ =hyperparameters).

L	OPERATION	COMPLEXITY	#
5	gradient step in full param. space	model dependent	$\times k_1$
6	transformation to symmetry space	$O(TNd^2)$	$\times 1$
8	nat. grad. step in symmetry space	$O(Td^3 + T^2d^2)$	$\times k_2$
9	transformation to full param. space	$O(TNd^2)$	$\times 1$

## 5. Experiments

We evaluate the proposed Goldstone-GD optimization algorithm on the three example models introduced in Section 3.2. We compare Goldstone-GD to standard GD, to AdaGrad (Duchi et al., 2011), and to Adam (Kingma & Ba, 2014). Goldstone-GD converges orders of magnitude faster and fits more interpretable word embeddings.

### 5.1. Visualizing Goldstone Modes With Artificial Data

**Model and Data Preparation.** We fit the dynamic Gaussian matrix factorization model defined in Eqs. 4-6 in Section 3.2 to small scale artificial data. In order to visualize Goldstone modes in the embedding space we choose an embedding dimension of  $d = 3$  and, for this experiment only, we fit the model to time independent-data. This allows us to monitor convergence since we know that the matrices  $U_t^*$  and  $V_t^*$  that minimize the loss are also time-independent. We generate artificial data for the matrix  $X \in \mathbb{R}^{10 \times 10}$  by drawing the components of two matrices  $\bar{U}, \bar{V} \in \mathbb{R}^{3 \times 10}$  from a standard normal distribution, forming  $\bar{U}^\top \bar{V}$ , and adding uncorrelated Gaussian noise with variance  $10^{-3}$ . We use  $T = 30$  time steps and a coupling strength of  $\lambda = 10$ .

**Hyperparameters.** We train the model with standard GD (baseline) and with Goldstone-GD with  $k_1 = 50$  and  $k_2 = 10$ . We find fastest convergence for the baseline method if we clip the gradients to an interval  $[-\bar{g}, \bar{g}]$  and use a decreasing learning rate  $\rho_s = \rho_0(\bar{s}/(s + \bar{s}))^{0.7}$  despite the noise-free gradient. Here,  $s$  is the training step. We optimize the hyperparameters for fastest convergence in the baseline and find  $\bar{g} = 0.01$ ,  $\rho_0 = 1$ , and  $\bar{s} = 100$ .

**Results.** Figure 2 compares convergence in the two algorithms. We discussed the figure at the end of Section 3.3. In summary, Goldstone-GD converges an order of magnitude faster even in this small scale setup that allows only for three different kinds of Goldstone modes (the skew symmetric gauge fields  $\Gamma_t$  have only  $d(d-1)/2 = 3$  independent parameters). Once the minimization finds minima of the local losses  $\ell$ , differences in the total loss  $\mathcal{L}$  between the two algorithms are small since Goldstone modes contribute only little to  $\mathcal{L}$  (this is why they decay slowly in GD).

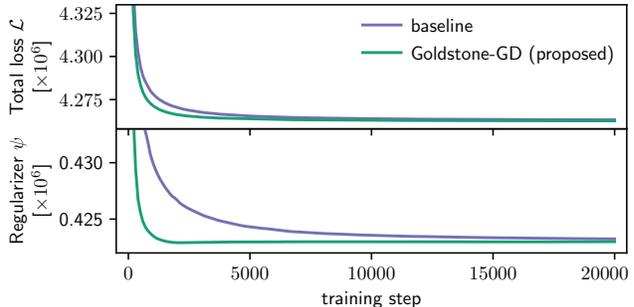


Figure 3. Training curves for MovieLens recommendations using sparse dynamic matrix factorization. Curves for three different random initializations are indistinguishable at this scale. Horizontal axis not accounting for a 1% runtime overhead in Goldstone-GD.

### 5.2. MovieLens Recommendations

**Model and Data Set.** We fit the sparse dynamic Bernoulli factorization model defined in Eqs. 6-9 in Section 3.2 to the MovieLens 20M data set<sup>2</sup> (Harper & Konstan, 2016). We use embedding dimension  $d = 30$ , coupling strength  $\lambda = 10$ , and regularizer  $\gamma = 1$ . The data set consists of 20 million reviews of 27,000 movies by 138,000 users with time stamps from 1995 to 2015. We binarize the ratings by splitting at the median, discarding ratings at the median, and we slice the remaining 18 million data points into  $T = 100$  time bins of equal duration. We split randomly across all bins into 50% training, 20% validation, and 30% test set.

**Baseline and Hyperparameters.** We compare the proposed Goldstone-GD algorithm to GD with AdaGrad (Duchi et al., 2011) with a learning rate prefactor of 1 obtained from cross-validation. Similar to Goldstone-GD, AdaGrad is designed to escape shallow valleys of the loss, but it uses only diagonal preconditioning. We compare to Goldstone-GD with  $k_1 = 100$  and  $k_2 = 10$ , using the same AdaGrad optimizer for update steps in the full parameter space.

**Results.** The additional operations in Goldstone-GD lead to a 1% overhead in runtime. The upper panel in Figure 3 shows training curves for the loss  $\mathcal{L}$  using the baseline (purple) and Goldstone-GD (green). The loss  $\mathcal{L}$  drops faster in Goldstone-GD, but differences in terms of the full loss  $\mathcal{L}$  are small because the local loss functions  $\ell$  are much larger than the regularizer  $\psi$  in this experiment. The lower panel of Figure 3 shows only  $\psi$ . Both algorithms converge to the same value of  $\psi$ , but Goldstone-GD converges at least an order of magnitude faster. The difference in value is small because Goldstone modes contribute little to  $\psi$ . They can, however, have a large influence on the parameter values, as we show next in experiments with Dynamic Word Embeddings.

<sup>2</sup><https://grouplens.org/datasets/movielens/20m/>

Table 2. Word aging: We translate modern words to the year 1800 using the shared representation space of Dynamic Word Embeddings.

QUERY	GOLDSTONE-GD	BASELINE
car	boat, saddle, canoe, wagon, box	shell, roof, ceiling, choir, central
computer	perspective, telescope, needle, mathematical, camera	organism, disturbing, sexual, rendering, bad
DNA	potassium, chemical, sodium, molecules, displacement	operates, differs, sharing, takes, keeps
electricity	vapor, virus, friction, fluid, molecular	exercising, inherent, seeks, takes, protect
tuberculosis	chronic, paralysis, irritation, disease, vomiting	trained, uniformly, extinguished, emerged, widely

### 5.3. Dynamic Word Embeddings

**Model and Data Set.** We perform variational inference (Ranganath et al., 2014) in Dynamic Word Embeddings (DWE), see Section 3.2. We fit the model to digitized books from the years 1800 to 2008 in the Google Books corpus<sup>3</sup> (Michel et al., 2011) (approximately  $10^{10}$  words). We follow (Bamler & Mandt, 2017) for data preparation, resulting in a vocabulary size of 10,000, a training set of  $T = 188$  time step, and a test set of 21 time steps. The DWE paper proposes two inference algorithms: filtering and smoothing. We use the smoothing algorithm, which has better predictive performance than filtering but suffers from Goldstone modes. We set the embedding dimension to  $d = 100$  due to hardware constraints and train for 10,000 steps using an Adam optimizer (Kingma & Ba, 2014) with a decaying prefactor of the adaptive learning rate,  $\rho_s = \rho_0(\bar{s}/(s + \bar{s}))^{0.7}$ , where  $s$  is the training step,  $\rho_0 = 0.1$ , and  $\bar{s} = 1000$ . We find that this leads to better convergence than a constant prefactor. All other hyperparameters are the same as in (Bamler & Mandt, 2017). We compare the baseline to Goldstone-GD using the same learning rate schedule and  $k_1 = k_2 = 10$ , which leads to an 8% runtime overhead.

**Results.** By eliminating Goldstone modes, Goldstone-GD makes word embeddings comparable across the time dimension of the model. We demonstrate this in Table 2, which shows the result of ‘aging’ modern words, i.e., translating them from modern English to the English language of the year 1800. For each query word  $i$ , we report the five words  $i'$  whose embedding vectors  $u_{i',1}$  at the first time step (year 1800) have largest overlap with the embedding vector  $u_{i,T}$  of the query word at the last time step (year 2008). Overlap is measured in cosine distance (normalized scalar product), between the means of  $u_{i,T}$  and  $u_{i',1}$  under the variational distribution.

Goldstone-GD finds words that are plausible for the year 1800 while still being related to the query (e.g., means of transportation in a query for ‘car’). By contrast, the baseline method fails to find plausible results. Figure 4 provides more insight into the failure of the baseline method. It shows histograms of the cosine distance between word embeddings

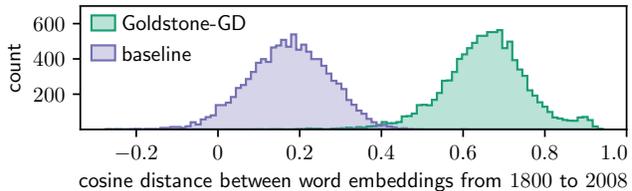


Figure 4. Cosine distance between word embeddings from the first and last year of the training data in Dynamic Word Embeddings.

$u_{i,1}$  and  $u_{i,T}$  for the same word  $i$  from the first to the last time step. In Goldstone-GD (green), most embeddings have a large overlap because the meaning of most words does not change drastically over time. By contrast, in the baseline (purple), no embeddings overlap by more than 60% between 1800 and 2008, and some embeddings even change their orientation (negative overlap). We explain this counterintuitive result with the presence of Goldstone modes, i.e., the entire embedding spaces are rotated against each other.

For a quantitative comparison, we evaluate the predictive log-likelihood of the test set under the posterior mean, and find slightly better predictive performance with Goldstone-GD ( $-0.5317$  vs.  $-0.5323$  per test point). The improvement is small because the training set is so large that the influence of the symmetry breaking regularizer is dwarfed in all but the symmetry directions by the log-likelihood of the data. The main advantage of Goldstone-GD are the more interpretable embeddings, as demonstrated in Table 2.

## 6. Conclusions

We identified a slow convergence problem in representation learning models with a continuous symmetry and a Markovian time series prior, and we solved the problem with a new optimization algorithm, Goldstone-GD. The algorithm separates the minimization in the symmetry subspace from the remaining coordinate directions. Our experiments showed that Goldstone-GD converges orders of magnitude faster and fits more interpretable embedding vectors, which can be compared across the time dimension of a model. Since continuous symmetries are common in representation learning, we believe that gauge theories and, more broadly, the theory of Lie groups are more widely useful in machine learning.

<sup>3</sup><http://storage.googleapis.com/books/ngrams/books/datasetsv2.html>

## Acknowledgements

We thank Ari Pakman for valuable and detailed feedback that greatly improved the manuscript.

## References

- Altland, A. and Simons, B. D. *Condensed matter field theory*. Cambridge University Press, 2010.
- Amari, S.-I. Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276, 1998.
- Badrinarayanan, V., Mishra, B., and Cipolla, R. Understanding symmetries in deep networks. *arXiv preprint arXiv:1511.01029*, 2015.
- Bamler, R. and Mandt, S. Dynamic word embeddings. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pp. 380–389, 2017.
- Barkan, O. Bayesian Neural Word Embedding. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 2017.
- Bui, H. H., Huynh, T. N., and Riedel, S. Automorphism groups of graphical models and lifted variational inference. *arXiv preprint arXiv:1207.4814*, 2012.
- Charlin, L., Ranganath, R., McInerney, J., and Blei, D. M. Dynamic Poisson factorization. In *Proceedings of the 9th ACM Conference on Recommender Systems*, pp. 155–162, 2015.
- Choi, S., Amari, S., Cichocki, A., and Liu, R.-w. Natural gradient learning with a nonholonomic constraint for blind deconvolution of multiple channels. In *First International Workshop on Independent Component Analysis and Signal Separation*, pp. 371–376, 1999.
- de Abreu, N. M. M. Old and new results on algebraic connectivity of graphs. *Linear Algebra and its Applications*, 423(1):53–73, 2007.
- Duchi, J., Hazan, E., and Singer, Y. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *Journal of Machine Learning Research*, 12: 2121–2159, 2011.
- Gens, R. and Domingos, P. M. Deep symmetry networks. In *Advances in Neural Information Processing Systems 27*, pp. 2537–2545. 2014.
- Gopalan, P., Hofman, J. M., and Blei, D. M. Scalable recommendation with hierarchical Poisson factorization. In *UAI*, pp. 326–335, 2015.
- Hall, B. *Lie groups, Lie algebras, and representations: an elementary introduction*, volume 222. Springer, 2015.
- Harper, F. M. and Konstan, J. A. The MovieLens Datasets: History and Context. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 5(4):19, 2016.
- Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. W. Stochastic Variational Inference. *Journal of Machine Learning Research*, 14(1):1303–1347, 2013.
- Kingma, D. and Ba, J. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference for Learning Representations (ICLR)*, 2014.
- Koren, Y. Collaborative filtering with temporal dynamics. *Communications of the ACM*, 53(4):89–97, 2010.
- LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- Lu, Z., Agarwal, D., and Dhillon, I. S. A spatio-temporal approach to collaborative filtering. In *ACM Conference on Recommender Systems (RecSys)*, 2009.
- Martens, J. New insights and perspectives on the natural gradient method. *arXiv preprint arXiv:1412.1193*, 2014.
- Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., et al. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, 331(6014): 176–182, 2011.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*, 2013a.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26*, pp. 3111–3119. 2013b.
- Mnih, A. and Salakhutdinov, R. R. Probabilistic matrix factorization. In *Advances in neural information processing systems*, pp. 1257–1264, 2008.
- Murphy, K. P. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.
- Noessner, J., Niepert, M., and Stuckenschmidt, H. Rockit: Exploiting parallelism and symmetry for map inference in statistical relational models. In *AAAI Workshop: Statistical Relational Artificial Intelligence*, 2013.
- Pascanu, R. and Bengio, Y. Revisiting natural gradient for deep networks. *arXiv preprint arXiv:1301.3584*, 2013.

- Pennington, J., Socher, R., and Manning, C. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.
- Poignard, C., Pereira, T., and Pade, J. P. Spectra of laplacian matrices of weighted graphs: structural genericity properties. *SIAM Journal on Applied Mathematics*, 78(1):372–394, 2018.
- Ranganath, R., Gerrish, S., and Blei, D. Black box variational inference. In *Artificial Intelligence and Statistics*, pp. 814–822, 2014.
- Rudolph, M. and Blei, D. Dynamic bernoulli embeddings for language evolution. *arXiv preprint arXiv:1703.08052*, 2017.
- Sun, F., Guo, J., Lan, Y., Xu, J., and Cheng, X. Sparse word embeddings using  $l_1$  regularized online learning. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pp. 2915–2921, 2016.
- Sun, J. Z., Varshney, K. R., and Subbian, K. Dynamic matrix factorization: A state space approach. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1897–1900, 2012.
- Vinyals, O. and Povey, D. Krylov subspace descent for deep learning. In *Artificial Intelligence and Statistics*, pp. 1261–1268, 2012.
- Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R. R., and Smola, A. J. Deep sets. In *Advances in Neural Information Processing Systems*, pp. 3394–3404, 2017.
- Zeiler, M. D. ADADELTA: an Adaptive Learning Rate Method. *arXiv preprint arXiv:1212.5701*, 2012.