A Quantum Field Theory of Representation Learning

Robert Bamler [1] **Stephan Mandt** [1]

Abstract

Continuous symmetries and their breaking play a prominent role in contemporary physics. Effective low-energy field theories around symmetry breaking states explain diverse phenomena such as superconductivity, magnetism, and the mass of nucleons. We show that such field theories can also be a useful tool in machine learning, in particular for loss functions with continuous symmetries that are spontaneously broken by random initializations. In this paper, we illuminate our earlier published work (Bamler & Mandt, 2018) on this topic more from the perspective of theoretical physics. We show that the analogies between superconductivity and symmetry breaking in temporal representation learning are rather deep, allowing us to formulate a gauge theory of 'charged' embedding vectors in time series models. We show that making the loss function gauge invariant speeds up convergence in such models.

1. Introduction

Continuous symmetries are of central interest in field theory. Gauge fields mediate interactions by propagating continuous symmetries transformations across space. Continuous symmetries are also common in representation learning, such as in word embeddings. In fact, the phrase 'dense embeddings' typically refers to models that do not distinguish any special directions in the representation space. Since no direction is special and only relative orientations matter, dense embedding models are often invariant under an arbitrary simultaneous rotations of all embedding vectors.

For example, factorizing a large matrix $X \approx U^\top V$ into two smaller matrices U and V is invariant under the transformation $U \mapsto RU, V \mapsto RV$ with any orthonormal matrix R since $(RU)^\top(RV) = U^\top R^\top R\, V = U^\top V$. In this paper, we consider models with continuous symmetries of this kind.

[1]Donald Bren School, Department of Computer Science, University of California at Irvine, USA. Correspondence to: Robert Bamler <rbamler@uci.edu>, Stephan Mandt <mandt@uci.edu>.

Presented at the ICML 2019 Workshop on Theoretical Physics for Deep Learning.

In (Bamler & Mandt, 2018), we recently showed that certain time series models with continuous symmetries suffer from slow convergence of Gradient Descent (GD). We proposed the Goldstone-GD algorithm, which speeds up convergence by using artificial gauge fields. While our motivation for Goldstone-GD came from the theory of superconductivity, the paper glossed over this connection as it addressed a general machine learning audience with no physics background.

Given the topic of this workshop, this paper primarily addresses readers that have some familiarity with concepts of theoretical physics, allowing them to understand these analogies on a deeper level. We expose a profound relation between the Goldstone-GD algorithm and gauge theory in general, and the theory of superconductivity in particular.

In Section 2, we discuss the concept of symmetry breaking in physics and in representation learning models for time series. The analogy shows that the considered models suffer from slow convergence due to a so-called gapless excitation spectrum, i.e., their loss function is ill-conditioned. In Section 3, we show that the analogy goes even further in that the considered models are similar to superfluids like helium-4 at low temperatures. This observation allows us to use a key insight: as *charged* superfluids (i.e., superconductors) have a gapped energy spectrum, we can solve the slow convergence problem by assigning a charge to embedding vectors, i.e., by coupling them to a gauge field. Section 4 repeats an experiment from (Bamler & Mandt, 2018).

2. Symmetry Breaking and Goldstone Modes

In this section, we introduce core concepts of the theory of spontaneous symmetry breaking in physics (Subsection 2.1) and apply them to machine learning (Subsection 2.2).

2.1. Goldstone Modes in Physics

We discuss the properties of Goldstone modes in physics. In Section 2.2, we show that the same theory also applies to Markovian time series models in machine learning.

Example: Phonons. Atoms in a solid arrange in a periodic lattice. This admits for a very compact description of the microscopic configuration: to provide the positions of all atoms, we only need to specify a global offset position with

respect to some reference lattice with the right periodicity. This phenomenon is called spontaneous symmetry breaking, and the global offset is an example of an order parameter. While the model (Hamiltonian) is invariant under arbitrary translations, the state breaks this continuous symmetry on a global level by picking a value for the order parmeter.

Real crystals are not perfectly periodic, and more realistic description therefore generalizes the order parameter to a smooth function of position and time. Waves in the order parameter field are called Phonons in crystals, and Goldstone modes in general. It turns out that Goldstone modes cost arbitrarily little energy in the long wavelength limit (Altland & Simons, 2010). For example, Figure 1a shows a measured Phonon dispersion relation, i.e., the phonon energy E as a function of the wave vector $q = \frac{2\pi}{\text{wave length}}$. We observe that $E \to 0$ for $q \to 0$, i.e., for smooth waves with long wavelength. One says that the phonon spectrum is 'gapless', i.e., the energy gap between the ground state and the lowest excitation goes to zero as the size of the system grows.

Other Examples of Goldstone Modes. Goldstone modes are as ubiquitous as spontaneous continuous symmetry breaking. For example, in a ferromagnet, the magnetization breaks rotational symmetry, and its Goldstone modes are called 'magnons'. In quantum chromo dynamics, 'pions' result from the spontaneous breaking of an approximate chiral symmetry. In superfluids, Goldstone modes arise due to spontaneous breaking of the $U(1)$ phase symmetry of quantum mechanics. As we discuss in Section 3.1, *charged* superfluids nevertheless have a gap in the energy spectrum due to the so-called Higgs mechanism. It is this mechanism that motivated the algorithm presented in this work.

2.2. Goldstone Modes in Time Series Models

We show that spontaneous continuous symmetry breaking arises in a general class of representation learning models for time series. We proof that the loss function of these models is ill-conditioned due to the existence of Goldstone modes. This subsection follows (Bamler & Mandt, 2018).

Representation Learning for Time Series. We introduce a general class of representation learning models for time series. Consider sequential data $\mathbf{X} \equiv \{X_t\}_{t=1:T}$ over T time steps $t \in \{1, \dots, T\}$. For each time step t, we fit a dense embedding model with parameters Z_t to the data X_t by minimizing some local loss function $\ell(X_t; Z_t)$ over Z_t. In our notation, the model parameters Z_t form a matrix whose columns are d-dimensional embedding vectors. As discussed in the introduction, we assume that the local loss function is invariant under arbitrary rotations R,

$$\ell(X_t; RZ_t) = \ell(X_t; Z_t) \qquad \forall R \in SO(d). \tag{1}$$

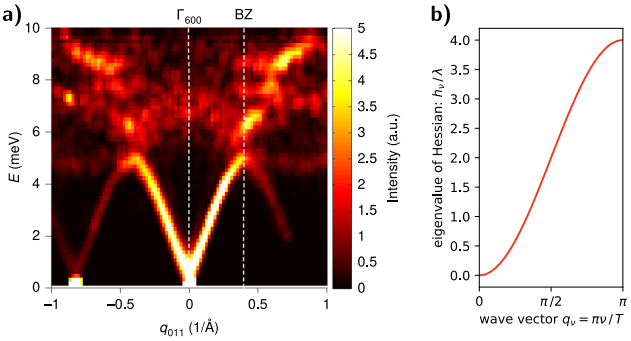

Figure 1. Dispersion relations of Goldstone modes; a) phonons in $Ba_{7.81}Ge_{40.67}Au_{5.33}$, measured with neutron scattering; b) eigenvalues h_ν of the Hessian of the regularizer in Eq. 2, see derivation in Section 2.2; both spectra are gapless, i.e., $E, h_\nu \to 0$ as $q \to 0$. This is a generic property of Goldstone modes. Figure a) taken from (Lory et al., 2017) with the authors' permission; CC-BY.

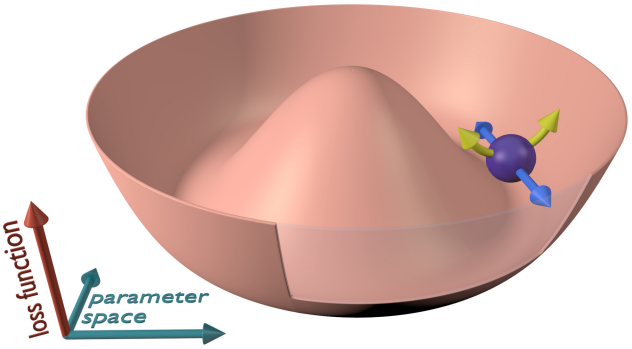

Figure 2. A rotationally symmetric potential ℓ has zero curvature within the symmetry subspace (blue arrows).

To share statistical strength across the time dimension, we add a quadratic Markovian (i.e., nearest neighbor) coupling with a strength λ. Denoting the concatenation of all model parameters as $\mathbf{Z} \equiv \{Z_t\}_{t=1:T}$, the total loss function is thus

$$\mathcal{L}(\mathbf{Z}) = \sum_{t=1}^{T} \ell(X_t; Z_t) + \frac{\lambda}{2} \sum_{t=2}^{T} ||Z_t - Z_{t-1}||_2^2. \tag{2}$$

Eqs. 1-2 define a general class of representation learning models for time series. The model class includes, e.g., dynamic matrix factorizations (Lu et al., 2009; Koren, 2010; Sun et al., 2012; Charlin et al., 2015) and dynamic word embeddings (Bamler & Mandt, 2017; Rudolph & Blei, 2018).

Goldstone Modes. We show that minimizing the loss function $\mathcal{L}$ defined in Eqs. 1-2 is an ill-conditioned optimization problem due to the existence of Goldstone modes. The argument goes in two steps. First, for each time step t, the local loss function ℓ has a manifold of degenerate minima: if Z_t^* minimizes $\ell(X_t; \cdot)$, then, according to Eq. 1, so

does RZ_t^* for any $R \in SO(d)$. The gradient $\nabla_{Z_t}\ell(X_t; Z_t)$ is zero and therefore constant over the entire manifold of degenerate minima, implying that the Hessian (i.e., the second derivative) has zero eigenvalues within the subspace spanned by infinitesimal local rotations (blue arrows in Figure 2). Thus, within this subspace, only the Hessian $\mathbf{H}_{\text{reg}}$ of the regularizer term in Eq. 2 contributes to the Hessian of $\mathcal{L}$.

Second, $\mathbf{H}_{\text{reg}}$ is ill-conditioned. The Markovian regularizer term in Eq. 2 has the form of the potential energy of a chain of T springs with spring constant λ. The eigenmodes of such a chain are harmonic waves $\propto \cos(q_\nu(t - \frac{1}{2}))$ with $q_\nu = \pi\nu/T$ for $\nu \in \{0, \dots, T-1\}$, and with eigenvalues $h_\nu = (2 - 2\cos(q_\nu))\lambda$, see Figure 1b. This is a *gapless* spectrum, i.e., in long time series (large T), the lowest eigenvalues become arbitrarily small: $h_\nu = O(\lambda/T^2)$ for small ν. This leads to a large condition number $h_{T-1}/h_1 = O(T^2)$ and therefore to slow convergence of gradient descent (GD).

3. Goldstone Gradient Descent

We propose a solution to the slow convergence problem from Section 2.2, motivated by an analogy to the quantum field theory of superconductivity. We discuss the physical intuition in Subsection 3.1 and apply it to time series models in Subsection 3.2. This work focuses on the fundamental physical principles invoked by the algorithm. A more pragmatic exposition with pseudocode instead of physical interpretations was given in (Bamler & Mandt, 2018).

3.1. Higgs Mechanism in Superconductivity

Our intuition for the algorithm described in Section 3.2 comes from comparing uncharged and charged superfluids. This section provides a very compact summary of these theories; for more details see, e.g., (Altland & Simons, 2010).

In quantum mechanics, the global phase of a wave function cannot be observed, implying a $U(1)$ symmetry. The ground state of an electrically neutral superfluid, e.g., helium-4 at low temperatures, breaks this symmetry spontaneously, leading to a Goldstone mode and a gapless excitation spectrum.

The situation is different in electrically charged superfluids, i.e., superconductors. For charged particles, the global $U(1)$ symmetry is promoted to a local gauge symmetry as the gradient of the wave function couples to the electromagnetic gauge fields (details see 'minimal coupling' in Section 3.2 below). The Goldstone mode inherits this coupling, leading to the Anderson-Higgs mechanism: the gauge field acquires a mass by 'consuming' the Goldstone mode. The photon mass makes magnetic fields energetically expensive so that the superconductor expels magnetic fields from its interior ('Meissner effect'). A quantitative derivation of the decay of magnetic fields at the surface of a superconductor requires considering the *dynamics* of the gauge fields.

3.2. Fast Optimization with Charged Embeddings

We propose an optimization method that quickly eliminates Goldstone modes, thereby solving the slow convergence problem discussed in Section 2.2. We follow three steps. First, using the physical intuition from Section 3.1, we give the embeddings $\mathbf{Z}$ a *charge*, i.e., we couple them to an auxiliary gauge field. Second, in analogy to our comments on the Meissner effect, we introduce a dynamics for the auxiliary gauge fields. Third, we apply a gauge transformation to turn the decay of gauge fields into a decay of Goldstone modes.

Minimal Coupling. Quantum mechanics encodes a state in a complex valued wave function $\psi(\mathbf{r}, t)$, where $\mathbf{r}$ and t are space and time coordinates, respectively. The theory is invariant under a global $U(1)$ symmetry that rotates the wave function by an arbitrary phase. However, it is not invariant under *local* phase rotations that map $\psi(\mathbf{r}, t)$ to $\tilde{\psi}(\mathbf{r}, t) := e^{i\varphi(\mathbf{r},t)}\psi(\mathbf{r}, t)$ with a local phase $\varphi(\mathbf{r}, t) \in \mathbb{R}$.

For electrically charged particles, the global $U(1)$ symmetry is promoted to a local gauge symmetry by coupling $\psi(\mathbf{r}, t)$ to so-called gauge fields $V(\mathbf{r}, t)$ and $\mathbf{A}(\mathbf{r}, t)$. The couplings are such that all observable effects of the above local phase rotation are compensated if we also change the gauge fields to $\tilde{V} := V - \nabla_t\varphi$ and $\tilde{\mathbf{A}} := \mathbf{A} + \nabla_{\mathbf{r}}\varphi$, respectively. This is called a gauge transformation.

We adapt the concept of gauge invariance to the time series models from Section 2.2. Due to Eq. 1 and the isotropic regularizer, the loss $\mathcal{L}$ in Eq. 2 is globally $SO(d)$ symmetric,

$$\mathcal{L}(RZ_1, \dots, RZ_T) = \mathcal{L}(Z_1, \dots, Z_T) \;\; \forall R \in SO(d). \quad (3)$$

We elevate this global symmetry to a local gauge symmetry by introducing t-dependent rotation matrices $R_t \in SO(d)$ $\forall t \in \{1, \dots, T\}$. Similar to how we expressed local phase rotations above as $e^{i\varphi(\mathbf{r},t)}$, we can enforce the constraint $R_t \in SO(d)$ by parameterizing R_t as the matrix exponential of a skew symmetric matrix $\Gamma_t = -\Gamma_t^\top$,

$$R_t := e^{\Gamma_t} = I + \Gamma_t + \frac{1}{2}\Gamma_t^2 + \frac{1}{3!}\Gamma_t^3 + \ldots \in SO(d) \quad (4)$$

Using the shorthand $\mathbf{\Gamma} \equiv \{\Gamma_t\}_{t=1:T}$, we define the gauge invariant loss function,

$$\mathcal{L}'(\mathbf{Z}; \mathbf{\Gamma}) := \mathcal{L}(R_1 Z_1, \dots, R_T Z_T). \quad (5)$$

By construction, $\mathcal{L}'$ is invariant under gauge transformations. Such a gauge transformation changes Γ_t to any other skew symmetric matrix $\tilde{\Gamma}_t$, and Z_t to $\tilde{Z}_t := e^{-\tilde{\Gamma}_t}e^{\Gamma_t}Z_t$. We thus say that the embeddings $\mathbf{Z}$ acquire a charge, and we call the components of $\mathbf{\Gamma}$ 'gauge fields'.[1] Note that $e^{-\tilde{\Gamma}_t}e^{\Gamma_t}$ can not be simplified to $e^{\Gamma_t - \tilde{\Gamma}_t}$ since $SO(d)$ is a non-abelian group. In this regard, the model is more similar to quantum chromodynamics than it is to quantum electrodynamics.

[1] Strictly speaking, gauge fields are finite differences of Γ_t's.

Minimization over Gauge Fields. The next step is to introduce a dynamics for the gauge fields $\mathbf{\Gamma}$. To find optimal rotations R_t, we minimize $\mathcal{L}'$ over $\mathbf{\Gamma}$ for fixed $\mathbf{Z}$. We thus insert Eqs. 4-5 into the loss function, Eq. 2. Because the local loss functions ℓ are rotationally symmetric (Eq. 1), they are independent of $\mathbf{\Gamma}$ and we only have to minimize the regularizer term. We truncate $\mathcal{L}'$ after the quadratic term in $\mathbf{\Gamma}$, resulting in the objective function

$$\sum_{t=2}^{T} \mathrm{Tr}\left[\left(\Gamma_t - \Gamma_{t-1} - \frac{1}{2}(\Gamma_t - \Gamma_{t-1})^2\right) Z_{t-1} Z_t^\top\right] \quad (6)$$

where the $d \times d$ matrices $Z_{t-1} Z_t^\top$ can be precalculated at the beginning of the optimization. The truncation after the quadratic term in $\mathbf{\Gamma}$ is asymptotically exact as, due to the gauge transformation described in the next paragraph, Goldstone modes decay over the course of the minimization and the minimum $\mathbf{\Gamma}^*$ of $\mathcal{L}'$ therefore approaches $\mathbf{\Gamma}^* \to \mathbf{0}$.

The minimization over $\mathbf{\Gamma}$ is in principle equally ill-conditioned as the original minimization problem. However, this is not an issue in practice. First, Eq. 6 is an explicit quadratic form in $\mathbf{\Gamma}$. We can therefore minimize it efficiently either by solving a system of linear equations, or, if this is too expensive, by using GD with full-rank preconditioning ('natural gradients'). Second, such specialized optimization methods are computationally possible because the optimization over $\mathbf{\Gamma}$ is over a much lower dimensional space than that over $\mathbf{Z}$: $\mathbf{\Gamma}$ lives entirely in the embedding space, i.e., unlike $\mathbf{Z}$, its matrix dimensions are independent of the number of embedding vectors.

Gauging Away Goldstone Modes. Minimizing over $\mathbf{\Gamma}$ reduces the gauge invariant loss function $\mathcal{L}'(\mathbf{Z}; \mathbf{\Gamma})$ but it leaves the embeddings $\mathbf{Z}$ and therefore the original loss $\mathcal{L}(\mathbf{Z})$ invariant. The final step of the algorithm is to apply a gauge transformation that reduces the value of $\mathcal{L}(\mathbf{Z})$ by eliminating Goldstone modes. We update the embedding vectors as $Z_t \leftarrow R_t Z_t = e^{\Gamma_t} Z_t$. As in the optimization phase, truncating the matrix exponential function to a finite order leads to an asymptotically correct result as $\mathbf{\Gamma} \to \mathbf{0}$.

Overall Training Loop. The gauge transformation minimizes $\mathcal{L}$ only along the directions of local rotations. Since the local loss functions ℓ are invariant under local rotations, they remain unchanged under gauge transformations. To optimize over the entire embedding space, we alternate between a phase of the above gauge field optimization, and a phase of standard gradient descent on $\mathcal{L}$ over $\mathbf{Z}$. This concludes the Goldstone-GD optimization algorithm.

4. Experiment

The focus of this paper is not the Goldstone-GD algorithm itself, but the physical intuition behind it. We therefore limit the discussion of experiments and present only one result from (Bamler & Mandt, 2018). We applied the proposed Goldstone-GD optimization algorithm to fit Dynamic Word Embeddings (Bamler & Mandt, 2017). The model combines $T = 188$ instances of a probabilistic version of word2vec (Mikolov et al., 2013; Barkan, 2017) with a time series prior similar to the regularizer in Eq. 2. We fitted the model to the Google Books corpus[2] (Michel et al., 2011), following the data preparation in (Bamler & Mandt, 2017).

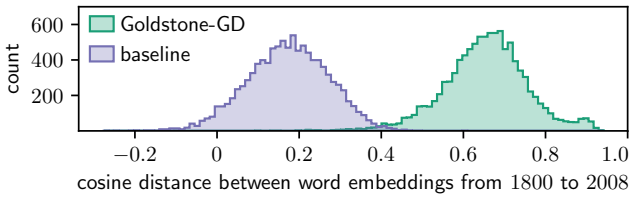

Figure 3. Cosine distance between word embeddings from the first and last year of the training data in Dynamic Word Embeddings.

We set ourselves a task that is very sensitive to the presence of Goldstone modes. Given a modern query word w, we want to find words w' that described the same concept in the year 1800. To this end, we look up the embedding $z_{w,2008}$ that the fitted model assigns to w in the year 2008 (the last year in the corpus). We then search among all embeddings in the year 1800 for the five embeddings $z_{w',1800}$ with largest cosine similarity (normalized scalar product) to $z_{w,2008}$.

Training the model without our advanced optimization algorithm lead to poor performance on this word aging task. For example, translating the word 'tuberculosis' (which was coined in 1839) from the year 2008 to the year 1800 resulted in the top five words 'trained', 'uniformly', 'extinguished', 'emerged', and 'widely', which seem unrelated to the query word. Our proposed Goldstone-GD algorithm, by contrast, provided reasonable results: 'chronic', 'paralysis', 'irritation', 'disease', and 'vomiting'. More examples of the word aging task can be found in (Bamler & Mandt, 2018).

Figure 3 provides a possible explanation for the failure of the baseline optimization method. It shows histograms for the cosine similarity between $z_{w,2008}$ and the corresponding vectors $z_{w,1800}$ for the same word w. In the baseline (purple), no word-embeddings overlap by more than 60%, in strong disagreement with our prior belief that only few words change their meaning over time. It suggests that the embedding spaces are misaligned, which is only weakly penalized if the two spaces are connected smoothly along the time axis, i.e., via a Goldstone mode. Our proposed method does not suffer from misaligned representation spaces because it eliminates Goldstone modes.

[2] http://storage.googleapis.com/books/ngrams/books/datasetsv2.html

References

Altland, A. and Simons, B. D. *Condensed Matter Field Theory*. Cambridge University Press, 2010.

Bamler, R. and Mandt, S. Dynamic Word Embeddings. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pp. 380–389, 2017.

Bamler, R. and Mandt, S. Improving Optimization in Models With Continuous Symmetry Breaking. In *International Conference on Machine Learning*, pp. 432–441, 2018.

Barkan, O. Bayesian Neural Word Embedding. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

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.

Koren, Y. Collaborative Filtering with Temporal Dynamics. *Communications of the ACM*, 53(4):89–97, 2010.

Lory, P.-F., Pailhès, S., Giordano, V. M., Euchner, H., Nguyen, H. D., Ramlau, R., Borrmann, H., Schmidt, M., Baitinger, M., Ikeda, M., et al. Direct measurement of individual phonon lifetimes in the clathrate compound $Ba_{7.81}Ge_{40.67}Au_{5.33}$. *Nature communications*, 8(1):491, 2017.

Lu, Z., Agarwal, D., and Dhillon, I. S. A Spatio–Temporal Approach to Collaborative Filtering. In *ACM Conference on Recommender Systems (RecSys)*, 2009.

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., 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. 2013.

Rudolph, M. and Blei, D. Dynamic Embeddings for Language Evolution. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pp. 1003–1011, 2018.

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.