Recurrently Predicting Hypergraphs

David W. Zhang
University of Amsterdam
w.d.zhang@uva.nl

Gertjan J. Burghouts
TNO
gertjan.burghouts@tno.nl

Cees G. M. Snoek
University of Amsterdam
cgmsnoek@uva.nl

Abstract

This work considers predicting the relational structure of a hypergraph for a given set of vertices, as common for applications in particle physics, biological systems and other complex combinatorial problems. A problem arises from the number of possible multi-way relationships, or hyperedges, scaling in $O(2^n)$ for a set of $n$ elements. Simply storing an indicator tensor for all relationships is already intractable for moderately sized $n$, prompting previous approaches to restrict the number of vertices a hyperedge connects. Instead, we propose a recurrent hypergraph neural network that predicts the incidence matrix by iteratively refining an initial guess of the solution. We leverage the property that most hypergraphs of interest are sparsely connected and reduce the memory requirement to $O(nk)$, where $k$ is the maximum number of positive edges, i.e., edges that actually exist. In order to counteract the linearly growing memory cost from training a lengthening sequence of refinement steps, we further propose an algorithm that applies backpropagation through time on randomly sampled subsequences. We empirically show that our method can match an increase in the intrinsic complexity without a performance decrease and demonstrate superior performance compared to state-of-the-art models.

1 Introduction

This work considers the task of inferring the relational structure for a given set of entities. In terms of a hypergraph the set of entities can be understood as the vertices, while the relations are expressed by the (hyper)edges. A hypergraph generalizes the concept of ordinary graphs to include edges that connect more than two vertices. This setting encompasses a wide spectrum of problems, like vertex reconstruction in particle physics [Shlomi et al., 2020a; Serviansky et al., 2020], inferring higher-order interactions in biological and social systems [Brugere et al., 2018; Battiston et al., 2020] or combinatorial optimization problems, such as finding the convex hull or Delaunay triangulation [Vinyals et al., 2015; Serviansky et al., 2020]. The number of possible edges grows in $O(2^n)$ with the number of vertices $n$, further implying a super-exponential growth for the number of possible hypergraphs. Simply just storing a single hypergraph in memory becomes challenging already for moderately sized $n$, as observed for Set2Graph by Serviansky et al. [2020], not to mention searching for the correct hypergraph. In this work we assume the hypergraphs of interest only contain a small fraction of all possible edges, as is the case for many relevant tasks. We show it is unnecessary to explicitly represent all non-existent edges and demonstrate the feasibility of solely supervising and predicting the positive edges.

Combinatorial optimization problems introduce the additional challenge of complexity to the learning task. Optimal solutions commonly require loops or recurrence, in order to repeat an operation multiple times, dependent on the input size [Grötschel et al., 2012]. A solution for finding the convex hull of a $d$-dimensional point set is optimal in $O(n \log(n) + n^{d/2})$ time [Chazelle, 1993]. Neural networks running on Turing machine equivalents are unable to circumvent the computational complexity intrinsic to a problem [Arora and Barak, 2009], albeit some operations may be parallelizable. Many known optimal solutions to computationally challenging problems are believed to be inherently

Preprint. Under review.
We represent vertices and edges as \( d \)-dimensional and \( E \)-dimensional vectors. Each entry \( I_{i,j} \) in the incidence matrix indicates if an edge \( e_i \) is incident to vertex \( v_j \). Notably, in our definition, an edge is not defined as a subset of vertices. Instead, the relational structure of the hypergraph is solely expressed by \( I \). We assume the number of edges is small, relative to the number of potential candidates, which is the case in many applications as we show in the experiments. During training, we leverage this property by setting \( k \), the maximum number of edges for any training example and effectively reduce the memory requirement of \( I \) to a tractable size. Furthermore, we relax the constraint on \( I_{i,j} \in \{0, 1\} \) to continuous values \( I_{i,j} \in [0, 1] \), in order to facilitate the usage of standard neural networks for modelling the incidence structure. The continuous incidence matrix instead indicates the probability that \( e_i \) and \( v_j \) are incident.

Initialization. \( \mathcal{V} \) is initialized from a learned affine transformation of the set of input entities. Each element in \( \mathcal{E} \) is sampled i.i.d. from an isotropic normal distribution with learned parameters. For \( I \), we do not assume the existence of any connections at the beginning and correspondingly initialize it to 0. It is important that the elements in \( \mathcal{E} \) are independently sampled from a distribution with sufficient variance. For any two edges in \( \mathcal{E} \) with identical feature vectors, any permutation equivariant neural network will map both elements to the same output vector. In practice, this property can already become a problem when edges are sufficiently similar. We observe it is unnecessary to explicitly lower bound the variance for our chosen architecture. We denote the initial hypergraph as \( \mathcal{H}^{(0)} \).

Refinement step. The refinement function \( \phi \) performs a holistic update \( \phi(\mathcal{H}^{(t-1)}) = \mathcal{H}^{(t)} \) at each iteration \( t \). We instantiate \( \phi \) as three different neural networks \( \phi_V, \phi_E, \phi_I \), each responsible for updating the vertices, edges and the incidence matrix, respectively. Each entry \( I_{i,j} \) in the incidence matrix is updated independently at each iteration \( t \), based on the \( j \)-th vertex, the \( i \)-th edge and the incidence from the previous step \( t-1 \):

\[
I^{(t)}_{i,j} = \phi_I \left( v^{(t-1)}_j, e^{(t-1)}_i, I^{(t-1)}_{i,j} \right),
\]

Figure 1: Recurrently predicting hypergraphs. The input set and an initial guess on the edges are combined into a hypergraph. A recurrent network holistically refines the hypergraph into the prediction. During training, we skip some iterations in the backward pass to improve efficiency and scale to a large number of iterations at constant memory cost.

2 Method

Definition. An undirected hypergraph \( \mathcal{H} \) is a 3-tuple \( \mathcal{H} = (\mathcal{V}, \mathcal{E}, I) \), where \( \mathcal{V} = \{v_1, \ldots, v_n\} \) is the set of vertices, \( \mathcal{E} = \{e_1, \ldots, e_k\} \) is the set of (hyper)edges and \( I \in \{0, 1\}^{k \times n} \) is the incidence matrix. We represent vertices and edges as \( d_V \)-dimensional and \( d_E \)-dimensional vectors. Each entry \( I_{i,j} \) in the incidence matrix indicates if an edge \( e_i \) is incident to vertex \( v_j \). Notably, in our definition, an edge is not defined as a subset of vertices. Instead, the relational structure of the hypergraph is solely expressed by \( I \). We assume the number of edges is small, relative to the number of potential candidates, which is the case in many applications as we show in the experiments. During training, we leverage this property by setting \( k \) equal to the maximum number of edges for any training example and effectively reduce the memory requirement of \( I \) to a tractable size. Furthermore, we relax the constraint on \( I_{i,j} \in \{0, 1\} \) to continuous values \( I_{i,j} \in [0, 1] \), in order to facilitate the usage of standard neural networks for modelling the incidence structure. The continuous incidence matrix instead indicates the probability that \( e_i \) and \( v_j \) are incident.

Initialization. \( \mathcal{V} \) is initialized from a learned affine transformation of the set of input entities. Each element in \( \mathcal{E} \) is sampled i.i.d. from an isotropic normal distribution with learned parameters. For \( I \), we do not assume the existence of any connections at the beginning and correspondingly initialize it to 0. It is important that the elements in \( \mathcal{E} \) are independently sampled from a distribution with sufficient variance. For any two edges in \( \mathcal{E} \) with identical feature vectors, any permutation equivariant neural network will map both elements to the same output vector. In practice, this property can already become a problem when edges are sufficiently similar. We observe it is unnecessary to explicitly lower bound the variance for our chosen architecture. We denote the initial hypergraph as \( \mathcal{H}^{(0)} \).

Refinement step. The refinement function \( \phi \) performs a holistic update \( \phi(\mathcal{H}^{(t-1)}) = \mathcal{H}^{(t)} \) at each iteration \( t \). We instantiate \( \phi \) as three different neural networks \( \phi_V, \phi_E, \phi_I \), each responsible for updating the vertices, edges and the incidence matrix, respectively. Each entry \( I_{i,j} \) in the incidence matrix is updated independently at each iteration \( t \), based on the \( j \)-th vertex, the \( i \)-th edge and the incidence from the previous step \( t-1 \):
We set \( k \) which replaces the incidence information used for aggregation in different hypergraph instances may admit differing number of edges. \( F \) variable number of edges. We instantiate as DeepSets [Zaheer et al., 2017] in the experiments. The aggregation functions \( \rho_{E \rightarrow V} \), \( \rho_{V \rightarrow E} \) sum over all edges and vertices, weighted by their respective incidence probability at time \( t \). We add skip-connections to \( \phi_{V} \) and \( \phi_{E} \) for \( V^{(t-1)} \) and \( E^{(t-1)} \), to facilitate training large number of iterations [He et al., 2016]. Additionally, we apply layer normalization [Ba et al., 2016] to both the input and output (following the skip connection) of \( \phi_{V} \) and \( \phi_{E} \). Previous works observed both the skip-connections and layer normalization to help speed up the training convergence in the context of a recurrent neural network [Locatello et al., 2020]. While the concatenation facilitates cross-communication between vertices and edges, the set neural network directly supplies context from the other vertices and edges, avoiding a detour through the incidence structure.

**Variable number of edges.** Different hypergraph instances may admit differing number of edges. We set \( k \) equal to the cardinality of the largest set of edges amongst all training examples and pad the target incidence matrix with additional rows of 0, resulting in an incidence matrix of shape \( k \times n \) for all examples. An edge that is incident to no vertex, i.e., \( I_{e} = 0 \), can be interpreted as non-existent. Usually the values of the predicted incidence matrix will not be exactly 0 or 1, requiring an additional decision procedure to determine if an edge exists and works: that it connects to a simple approach is to decide the incidence structure locally for each entry in the incidence matrix. This approach ignores fixed structures that are inherent in many problems, e.g., each edge connects exactly \( 3 \) (or no) vertices in a convex hull of \( 3 \)-dimensional point sets. To improve the naive local version, while keeping the decision process simple, we add an existence indicator \( \sigma_{i}^{(t)} \) that is computed as a dot product between the edge \( e_{i}^{(t)} \) and a learned weight vector, followed by a sigmoid function. Using the existence indicator, we update the incidence matrix:

\[
I_{i}^{(t)} = \sigma_{i}^{(t)} I_{i}^{(t)},
\]

which replaces the incidence information used for aggregation in \( \rho_{E \rightarrow V} \) and \( \rho_{V \rightarrow E} \). With an explicit existence indicator we can simply choose the \( 3 \) vertices with the highest incidence probability for each existing edge in the 3D convex hull example.

**Iterative refinement.** The prediction at test-time is attained after applying the refinement step, described in Equations (1) to (4), for \( T_{\text{total}} \) number of times. Our approach can be considered an instance of a recurrent neural network (RNN), as each refinement step uses the same set of parameters at each iteration \( t \). Different from conventional RNNs, we structure the latent state of our network as a hypergraph and inject the same input at every time step. Intuitively, the recurrent architecture provides the neural network with multiple processing steps to consider the global context \( I \), in the prediction of the local structures \( I_{i,j} \). In contrast to auto-regressive approaches, we consider all \( I_{i,j} \) simultaneously. Thereby, we do not impose any artificial order on the edges and facilitate parallelization in the problem-solving procedure.

**BPTT with gradient skips.** The common learning method for RNNs is backpropagation through time (BPTT) [Pearlmutter, 1995; Robinson and Fallsida, 1987; Werbos, 1988; Mozer, 1989]. Conceptually BPTT treats the forward process as a feedforward neural network with weight-tied parameters and applies backpropagation to update the network parameters with respect to the gradients of a loss function. Applying BPTT on the entire sequence of \( T_{\text{total}} \) refinement steps is prohibitive for large \( T_{\text{total}} \), due to memory costs growing in \( O(T_{\text{total}}) \). Instead, we approximate the complete gradient by applying BPTT on selected subsequences of length \( T_{\text{BPTT}} \), with skips in between different subsequences. By limiting the length of the backward pass in this way, our memory usage remains constant over time. We select subsequences such that there are zero or more iterations in between any pair of subsequences. In the case when there are exactly \( 0 \) iterations in between any consecutive subsequences, we recover truncated BPTT [Williams and Peng, 1990], which in its more general form
We apply the loss on each intermediate incidence matrix. We match the edges between the target and the prediction with Hungarian matching [Kuhn, 1955]. (edges) may be in arbitrary order due to independent and identically distributed random initialization.

**Training objective.** During training, we minimize the element-wise binary cross entropy of the target and predicted incidence matrix. While the column (vertices) order is given by the input, the rows (edges) may be in arbitrary order due to independent and identically distributed random initialization. We match the edges between the target and the prediction with Hungarian matching [Kuhn, 1955]. We apply the loss on each intermediate incidence matrix \( I^{(t)} \) to encourage iterative improvements.

**Computational complexity.** The space complexity of \( \mathcal{H} \) is in \( O(nk) \), offering an efficient representation for hypergraphs when the maximal number of edges \( k \) is relatively low. Problems that involve edges connecting many vertices benefit from this choice of representation, as the space requirement does not scale with the maximal connectivity of an edge. During training, BPTT further scales the memory requirement by a factor of \( T_{BPTT} \), possibly limiting the number of refinement steps. Our proposed training algorithm does not directly scale with \( T_{total} \), offering a tractable solution for problems requiring many iterations. The loss computation is dominated by the Hungarian matching, requiring time in \( O(k^3) \). This computational cost is further scaled by the factors \( N_{BPTT} \), \( T_{BPTT} \) and the mini-batch size. In practice, the impact from \( T_{BPTT} \) and the mini-batch size are dampened by applying the matching algorithm in parallel. In contrast to that, each gradient update step happens sequentially, motivating smaller choices for \( N_{BPTT} \), to reduce the training time. By randomly skipping refinement steps during training, we require less gradient update steps for fixed \( T_{BPTT} \) and \( T_{total} \), resulting in faster training.

### 3 Experiments

Next, we evaluate our approach on multiple set-to-hypergraph tasks, in order to assess our approach and examine the main design choices. For additional details, we refer to the appendix. Code is available at: [https://github.com/davzha/recurrently_predicting_hypergraphs](https://github.com/davzha/recurrently_predicting_hypergraphs).

#### 3.1 Does improving the asymptotic memory requirement degrade predictive performance?

The improved asymptotic memory cost of our method, hinges on the feasibility to train with only (or mainly) positive edge examples, as opposed to explicitly supervising negative edges. To assess if the computational efficiency improvement comes at the cost of decreased test-time performance, we compare our method to prior state-of-the-art on three common benchmarks [Serviansky et al., 2020].

**Particle partitioning.** Particle colliders are an important tool for studying the fundamental particles of nature and their interactions. During a collision, several particles are emanated and measured by nearby detectors, while some particles decay beforehand. Identifying which measured particles share a common progenitor is an important subtask in the context of vertex reconstruction [Shlomi et al., 2020]. For more information, we refer to [Shlomi et al., 2020a]. We use the simulated dataset of \( 0.9M \) examples (CC-BY 4.0) [Serviansky et al., 2020; Shlomi et al., 2020], with the same

---

**Algorithm 1: BPTT with skips**

Input: \( V_{\text{input}}, I_{\text{target}}, T_{\text{total}}, T_{\text{BPTT}}, N_{\text{BPTT}} \)

\( \mathcal{H} := (V, E, I) \leftarrow \text{initialize}(V_{\text{input}}) \)

\( T_{\text{skip}} \leftarrow \text{sample}(T_{\text{total}}, T_{\text{BPTT}}, N_{\text{BPTT}}) \)

for \( i \) in \( T_{\text{skip}} \):

| with no_grad():
| ---
| for \( i \) in range(\( t \)):
| \( \mathcal{H} \leftarrow \text{refine}(\mathcal{H}) \)
| \( \mathcal{L} \leftarrow 0 \)
| for \( i \) in range(\( T_{\text{BPTT}} \)):
| \( \mathcal{H} \leftarrow \text{refine}(\mathcal{H}) \)
| \( \mathcal{L} \leftarrow \mathcal{L} + \text{loss}(\mathcal{H}, I_{\text{target}}) \)
| \text{backward}(\mathcal{L})
| \text{gradient_step_and_reset()}
train/validation/test split. Each example is sampled from one of three different generation processes that differ in the typical number of partitions per set: bottom, charm and light. Each set of particles, also called jets, consists of 2 to 14 elements represented by 10 features. The target partitionings indicate the common progenitors and restrict the valid incidence matrices to those with a single incident edge per vertex.

We compare our method to Set2Graph [Serviansky et al., 2020], Set Transformer [Lee et al., 2019] and Slot Attention [Locatello et al., 2020]. Set2Graph inures a prohibitively large memory cost when predicting edges with high connectivity and instead resorts to only predicting edges with at most 2 connecting vertices, followed by an additional heuristic to infer the partitions. Both Set Transformer and Slot Attention were not designed for set-to-hypergraph tasks, but we consider them as general set-to-set models. Building on the idea of solely predicting the set of positive edges, we construct two additional baselines that map from the set of vertices to the set of edges, followed by an incidence indicator network similar to Equation 1. We train our method with $T_{\text{total}}=16$, $T_{\text{BPTT}}=4$, $N_{\text{BPTT}}=2$ and a maximum of 10 edges. Since each particle can only be part of a single partition, we choose the one with the highest incidence probability at test time. All models additionally minimize a soft $F_1$ score [Serviansky et al., 2020]. In Table 1 we report the performances on each type of jets as the $F_1$ score and Adjusted Rand Index (ARI). Our method outperforms all alternatives on bottom and charm jets, while being competitive on light jets. Coupled with the performance of the set-to-set based baselines, the results highlight that it is indeed feasible to train without explicitly supervising negative edges. Furthermore, the relative improvement of our approach compared to the set-to-set baselines hints at the benefits of a holistic refining approach.

**Convex hull.** The convex hull of a finite set of $d$-dimensional points can be efficiently represented as the set of simplices that enclose all points. In the 3D case, each simplex consists of 3 points that together form a triangle. For the general $d$-dimensional case, the valid incidence matrices are limited to those with $d$ incident vertices per edge. Finding the convex hull is an important and well understood task in computational geometry, with optimal exact solutions [Chazelle 1993, Preparata and Shamos 2012]. Nonetheless, predicting the convex hull for a given set of points poses a challenging problem for current machine learning methods, especially when the number of points increases [Vinyals et al., 2015, Serviansky et al., 2020]. Each point set is drawn from one of two different distributions: Gaussian and spherical. In the Gaussian case, each point is drawn i.i.d. from a standard normal distribution. In the spherical case we additionally normalize each point to lie on the unit sphere. In subsequent comparisons to prior work, we consider the same point distributions, set sizes and training/validation/test data sizes as in [Serviansky et al., 2020].

On convex hull finding in 3D, we compare our method to the same baselines as in the previous experiment. Set2Graph [Serviansky et al., 2020] learns to map the set of 3D points to the 3rd order adjacency tensor. Since storing this tensor in memory is not feasible, they instead concentrate on a local version of the problem, which only considers the $k$-nearest neighbors for each point [Serviansky et al., 2020]. We train our method with $T_{\text{total}}=48$, $T_{\text{BPTT}}=4$, $N_{\text{BPTT}}=6$ and set $k$ equal to the highest number of triangles in the training data. At test time, a prediction admits an edge $e_i$ if its existence indicator $\sigma_i > 0.5$. Each edge is incident to the three vertices with the highest incidence probability.

In Table 2 we report the $F_1$ score for $m=30$, $n=50$ and $n\in[20...100]$ for both Gaussian and spherically distributed point sets. The results demonstrate that our method consistently outperforms all the baselines by a considerable margin. Interestingly, our method does not suffer from the same drastic performance decline as in Set2Graph, when increasing the set size from 30 to 50. The results for $n\in[20...100]$ with Gaussian distributed points shows a stark drop in performance for Set2Graph.

### Table 1: Particle partitioning results.

On three jet types performance measured in $F_1$ score and adjusted rand index (ARI). Our method outperforms the baselines on bottom and charm jets, while being competitive on light jets.

| Model      | bottom jets | charm jets | light jets |
|------------|-------------|------------|------------|
|            | $F_1$       | ARI        | $F_1$      | ARI        | $F_1$      | ARI        |
| Set2Graph  | 0.646 ± 0.002 | 0.491 ± 0.008 | 0.747 ± 0.002 | 0.457 ± 0.004 | 0.972 ± 0.002 | 0.931 ± 0.002 |
| Set Transformer | 0.630 ± 0.004 | 0.464 ± 0.007 | 0.747 ± 0.003 | 0.460 ± 0.007 | 0.970 ± 0.001 | 0.922 ± 0.002 |
| Slot Attention | 0.600 ± 0.012 | 0.411 ± 0.021 | 0.728 ± 0.008 | 0.429 ± 0.016 | 0.963 ± 0.002 | 0.895 ± 0.009 |
| Ours       | 0.679 ± 0.002 | 0.548 ± 0.003 | 0.763 ± 0.001 | 0.499 ± 0.002 | 0.972 ± 0.001 | 0.926 ± 0.002 |
The goal in this task is to predict the adjacency matrix of an ordinary graph — a graph consisting of $n$ points to the set of Delaunay edges, represented by the adjacency matrix. Two of the main design choices of our approach remain in this adaptation: Iterative refining and effectively reducing the incidence matrix to an adjacency matrix, since it is computed based above 1 back to 1 and setting the diagonal to 0. This detour through the incidence matrix is clearly unfavorable, as in this case the incidence matrix actually takes up more compute than the adjacency matrix. Instead of applying our method directly, we consider a simple adaptation of our approach to.

The problem of Delaunay triangulation is, similar to convex hull finding, a well studied problem in computational geometry and has exact solutions in $O(n \log(n))$ [Rebay, 1993]. We consider the same learning task and setup as Serviansky et al. [2020], who frame Delaunay triangulation as mapping from a set of 2D points to the set of Delaunay edges, represented by the adjacency matrix. The point sets of size $n$ are sampled uniformly from the unit square, with $n=50$ or $n \in [20..80]$.

The goal in this task is to predict the adjacency matrix of an ordinary graph — a graph consisting of edges that connect two vertices — where the number of edges are greater than the number of vertices. One could recover the adjacency matrix based on the matrix product of $T^T T$, by clipping all values above 1 back to 1 and setting the diagonal to 0. This detour through the incidence matrix is clearly unfavorable, as in this case the incidence matrix actually takes up more compute than the adjacency matrix. Instead of applying our method directly, we consider a simple adaptation of our approach to the graph setting. We replace the initial set of edges with the (smaller) set of vertices and apply the same vertex refinements (Equation 2) on both sets. This change results in $E = V$ for the prediction and effectively reduces the incidence matrix to an adjacency matrix, since it is computed based on all pairwise combinations of $E$ and $V$. We further replace the concatenation in $\phi_T$ (Equation 1) with a sum, to ensure that the predicted adjacency matrix is symmetric and represents an undirected graph. Two of the main design choices of our approach remain in this adaptation: Iterative refining of the complete graph with a recurrent neural network and BPTT with gradient skips. We train our model with $T_{\text{total}}=32$, $T_{\text{BPTT}}=4$ and $N_{\text{BPTT}}=4$. At test-time, an edge between two vertices exists if the adjacency value is greater than 0.5.

In Table 3 we report the accuracy, precision, recall and F1 score for Set2Graph [Serviansky et al., 2020] and our adapted method. Our method outperforms Set2Graph on all metrics. More importantly, our method does not see a large discrepancy in performance between set sizes. We attribute this to the recurrence of our iterative refinement scheme, which we study in more depth subsequently.

**3.2 What is the role of recurrence in addressing the intrinsic complexity of a problem?**

The intrinsic complexity of finding a convex hull for a $d$-dimensional set of $n$ points is in $O(n \log(n) + n^{\frac{d}{2}})$ [Chazelle, 1993]. This scaling behavior offers an interesting opportunity to study the effects of increasing (time) complexity on model performance. The intrinsic time
complexity of the convex hull problem signifies that any solution scales super-linearly with the input set size. Since our learned model is not an exact solution, the implications of the complexity become less clear. In order to assess the relevancy of the problem’s complexity for our approach, we examine the relation between the number of refining steps and increases in the intrinsic resource requirement. The following experiments are all performed with standard BPTT, in order to not introduce additional hyperparameters that may affect the conclusions.

First we examine the performance of our approach with 3 iterations, trained on increasing set sizes \( n \in [10 \ldots 50] \). In Figure 2 we observe a drop in performance with increasing set sizes when training with the same number of iterations. The observed (negative) correlation between the set size and the performance confirms a relationship between the computational complexity and the difficulty of the learning task. Next, we examine the performance for varying number of iterations and set sizes. We refer to the setting, where the number of iterations is 3 and set size \( n = 10 \) as the base case. All other set sizes and number of iterations are chosen such that the performance matches the base case as closely as possible. In Figure 2, we observe that the required number of iterations increases with the input set size. The results highlight that an increase in the number of iterations suffices in counteracting the performance drop. The number of refinement steps scales sub-linearly with the set size, different from the intrinsic time complexity. This highlights the property that our approach does not operate incrementally, but instead parallelizes the edge finding process. Notably, increasing the number of iterations does not introduce any additional parameters, as opposed to increasing the depth or width. The two experiments suggest that the intrinsic complexity of a problem manifests similarly for solutions containing errors as they do for exact solutions. An analogy can be made with an exact algorithm that incrementally constructs the convex hull, but is allocated insufficient computing time, in which case the result will have (more) errors. By replacing BPTT with truncated BPTT or BPTT with skips, our approach offers a scalable solution for settings with a fixed memory budget.

3.3 How does recurrence compare to stacking layers?

This section examines the role of recurrence in our approach, by comparing to a baseline that stacks multiple refinement layers, each with a separate set of parameters. We compare the models trained for 3 to 9 refinement steps. Both versions are trained by backpropagating through the entire sequence of refinement steps. For the stacked version, we only compute the loss at the final refinement step, as we observed that supervising all intermediate steps leads to worse performance. In Figure 3a, we show that both cases benefit from an increasing number of refinement steps. Adding more iterations beyond 6 only slightly improves the performance of the stacked model, while the recurrent version still benefits, leading to an absolute difference of 0.03 in F1 score for 9 iterations. The results indicate that recurrence constitutes a useful inductive bias on an intrinsically complex problem, especially when supervised over multiple steps. Next we train a recurrent and stacked model, with 15 iterations, until both achieve a similar validation performance, by stopping training early on the recurrent model. The results in Figure 3b show that the recurrent variant performs better when tested at larger set sizes than trained, indicating an improved generalization ability.

3.4 Can we actually learn tasks that encompass higher-order edges?

The particle partitioning experiment exemplifies a case where a single edge can connect up to 14 vertices. Set2Graph [Serviansky et al., 2020] demonstrates that in this specific case it is possible to approximate the hypergraph with a graph. They leverage the fact that any vertex is incident to exactly one edge and apply a post-processing step that constructs the edges from noisy cliques. Instead, we consider a task for which no straightforward graph based approximation exists. Specifically, we consider convex hull finding in 10-dimensional space for 13 standard normal distributed points. We train with \( T_{\text{total}} = 32, N_{\text{BPTT}} = 4 \) and \( T_{\text{BPTT}} = 4 \). The test performance reaches an F1 score of 0.75, clearly demonstrating that the model managed to learn. This result demonstrates the improved
scaling behavior can be leveraged for tasks that are computationally out of reach for adjacency-based approaches.

3.5 Can we skip gradients without degrading the performance?

In order to assess if skipping gradients comes at the cost of decreased test-time performance, we compare our approach to truncated BPTT [Williams and Peng 1990]. We constrict both truncated BPTT and BPTT with skips to a fixed memory budget, by limiting any backward pass to the most recent $T_{\text{BPTT}}=4$ iterations, for $T_{\text{total}}\in\{16,32\}$. We consider two variants of our training algorithm: 1. Skipping iterations at fixed time steps and 2. Skipping randomly sampled time steps. In both the fixed and random skips versions, we skip half of the total iterations. We train all models on convex hull finding in 3-dimensions for 30 spherically distributed points. In addition, we include baselines trained with standard BPTT that contingently inform us about performance degradation incurred by truncating or skipping gradients. Standard BPTT increases the memory footprint linearly with the number of iterations $T_{\text{total}}$, inevitably exceeding the available memory at some threshold. We deliberately choose a small set size in order to accommodate training with standard BPTT for $T_{\text{total}}\in\{4,16,32\}$ number of iterations.

The results in Figure 4 demonstrate that skipping half of the iterations during backpropagation, significantly decreases the training time without incurring any test-time performance degradation. This suggests that supervising all iterations, like in truncated BPTT, is highly redundant and skipping iterations in the weight update phase constitutes an effective approach to reducing said redundancy. In general, we observe that sampling random subsequences to apply BPTT on is more stable than picking fixed time steps. Both truncated BPTT and BPTT with skips greatly outperform standard BPTT, when the memory budget is constricted to $T_{\text{BPTT}}=4$. The results on standard BPTT appear to indicate that performance worsens when increasing the number of iterations from 16 to 32. We observe that applying BPTT on many iterations leads to increasing gradient norms in the course of training, complicating the training process. The memory limited versions did not exhibit a similar behavior, evident from the improved performance, when increasing the iterations from 16 to 32.

4 Related Work

**Adjacency prediction.** Set2Graph [Serviansky et al., 2020] is a family of maximally expressive permutation equivariant neural networks that map from an input set to (hyper)graphs. The authors
extend the general idea, of applying a scalar-valued adjacency indicator function on all pairs of vertices [Kipf and Welling, 2016], to the $l$-edge case [Serviansky et al., 2020], i.e., edges that connect $l$ vertices. In Set2Graph, for each $l$ the adjacency structure is modeled by an $l$-tensor, requiring memory in $O(n^l)$. This becomes intractable already for small $l$ and moderate set sizes. By pruning the negative edges, our approach scales in $O(nk)$, making it applicable even when $l=n$.

Set prediction. Recent works on set prediction map a learned initial set [Zhang et al., 2019, Lee et al., 2019] or a randomly initialized set [Locatello et al., 2020, Carion et al., 2020, Zhang et al., 2021] to the target space. Out of these, the closest one to our hypergraph refining approach is Slot Attention [Locatello et al., 2020], which recurrently applies the Sinkhorn operator [Adams and Zemel, 2011] in order to associate each element in the input set with a single slot (hyperedge). None of the prior works on set prediction consider the set-to-hypergraph task, but some can be naturally extended by mapping the input set to the set of positive edges, an approach similar to ours.

Non-temporal recurrent networks. Although recurrence in the context of neural networks is commonly used to handle time-varying inputs [Elman, 1990] and outputs [Graves et al., 2006], many have applied recurrent networks in non-temporal settings [Hopfield, 1982, Locatello et al., 2020]. Deep implicit models [Ghaoui et al., 2019, Bai et al., 2019, 2020, Gu et al., 2020] recurrently apply a nonlinear transformation on the latent state until reaching a fixed point [Pineda, 1987, Simard et al., 1988]. These models’ memory consumption does not scale with the recurrence, by analytically deriving the gradient at the fixed point via the implicit function theorem [Krantz and Parks, 2012]. This comes at the cost of restricting the model to those that reliably converge to a fixed point [Bai et al., 2019]. While some works enforce the existence of a unique attracting fixed point [Winston and Kolter, 2020, Gu et al., 2020], this property is undesirable for our case. Any permutation equivariant neural network with a unique attracting fixed point will collapse its hidden state to a multiset of identical elements, making this approach inapplicable to set-to-hypergraph, without forgoing permutation equivariance. Our approach offers an alternative that retains the constant memory scaling, but does not necessitate any restrictions on the RNN.

Hypergraph neural network. Feng et al. [2019] propose an extension of graph convolutions [Defferrard et al., 2016, Kipf and Welling, 2017] to hypergraphs and apply it to classification tasks. Bai et al. [2021] apply the attention mechanism to induce incidence matrices, over which they deploy a hypergraph convolution operator. Our recurrent hypergraph neural network holistically refines the whole hypergraph at each iteration, updating vertex/edge features and the incidence structure. Furthermore, the hypergraph convolution [Bai et al., 2021] does not benefit from increased depth, as commonly observed for deep graph neural networks [Oono and Suzuki, 2019, Li et al., 2018, Nt and Maehara, 2019], where increased depth leads to over-smoothing. In experiments, we have demonstrated our recurrent network to greatly benefit from additional depth.

5 Limitations and Future Work

The scope of our work is limited to undirected hypergraphs, though extending this approach to directed hypergraphs constitutes an interesting future direction. Directed hypergraphs can represent targeted relations in applications like scene graph generation [Chang et al., 2021]. We identify the Hungarian matching as the main computational bottleneck during training. Replacing the Hungarian matched loss with a faster alternative, like a learned energy function [Zhang et al., 2021], would greatly speed up training for tasks with large $k$. In application areas such as social networks, interactions are typically modeled by graphs with pair-wise edges [Zhou et al., 2020]. Extending these to higher-order edges may facilitate new applications and opportunities, possibly invoking complex ethical and societal questions.

6 Conclusion

We have introduced a new hypergraph RNN, accompanied by a recurrent training algorithm for set-to-hypergraph prediction. By representing and supervising the set of positive edges, we substantially improve the asymptotic scaling and enable learning tasks with higher-order edges. On common benchmarks, we have demonstrated that our method outperforms previous works, while offering a more favorable asymptotic scaling behavior. In further evaluations, we have highlighted the importance of recurrence for addressing the intrinsic complexity of problems.
Acknowledgments and Disclosure of Funding

This work is part of the research programme Perspectief EDL with project number P16-25 project 3, which is financed by the Dutch Research Council (NWO) domain Applied and Engineering Sciences (TTW).

References

R. P. Adams and R. S. Zemel. Ranking via sinkhorn propagation. *arXiv preprint arXiv:1106.1925*, 2011.

S. Arora and B. Barak. *Computational complexity: a modern approach*. Cambridge University Press, 2009.

J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

S. Bai, J. Z. Kolter, and V. Koltun. Deep equilibrium models. *Advances in Neural Information Processing Systems*, pages 690–701, 2019.

S. Bai, V. Koltun, and J. Z. Kolter. Multiscale deep equilibrium models. In *Advances in Neural Information Processing Systems*, 2020.

S. Bai, F. Zhang, and P. H. Torr. Hypergraph convolution and hypergraph attention. *Pattern Recognition*, 110:107637, 2021.

F. Battiston, G. Cencetti, I. Iacopini, V. Latora, M. Lucas, A. Patania, J.-G. Young, and G. Petri. Networks beyond pairwise interactions: structure and dynamics. *Physics Reports*, 2020.

I. Brugere, B. Gallagher, and T. Y. Berger-Wolf. Network structure inference, a survey: Motivations, methods, and applications. *ACM Computing Surveys (CSUR)*, 51(2):1–39, 2018.

N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.

X. Chang, P. Ren, P. Xu, Z. Li, X. Chen, and A. Hauptmann. Scene graphs: A survey of generations and applications. *arXiv preprint arXiv:2104.01111*, 2021.

B. Chazelle. An optimal convex hull algorithm in any fixed dimension. *Discrete & Computational Geometry*, 10(4):377–409, 1993.

M. Defferrard, X. Bresson, and P. Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in Neural Information Processing Systems*, 2016.

J. L. Elman. Finding structure in time. *Cognitive Science*, 14(2):179–211, 1990.

Y. Feng, H. You, Z. Zhang, R. Ji, and Y. Gao. Hypergraph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3558–3565, 2019.

L. E. Ghaoui, F. Gu, B. Travacca, A. Askari, and A. Y. Tsai. Implicit deep learning. *arXiv preprint arXiv:1908.06315*, 2019.

A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *International Conference on Machine Learning*, pages 369–376, 2006.

R. Greenlaw, H. J. Hoover, W. L. Ruzzo, et al. *Limits to parallel computation: P-completeness theory*. Oxford University Press on Demand, 1995.

M. Grötschel, L. Lovász, and A. Schrijver. *Geometric algorithms and combinatorial optimization*, volume 2. Springer Science & Business Media, 2012.

F. Gu, H. Chang, W. Zhu, S. Sojoudi, and L. E. Ghaoui. Implicit graph neural networks. In *Advances in Neural Information Processing Systems*, pages 11984–11995, 2020.
K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In IEEE Conference on Computer Vision and Pattern Recognition, pages 770–778, 2016.

J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. Proceedings of the National Academy of Sciences, 79(8):2554–2558, 1982.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

T. N. Kipf and M. Welling. Variational graph auto-encoders. arXiv preprint arXiv:1611.07308, 2016.

T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. In International Conference on Learning Representations, 2017.

S. G. Krantz and H. R. Parks. The implicit function theorem: history, theory, and applications. Springer Science & Business Media, 2012.

H. W. Kuhn. The Hungarian method for the assignment problem. Naval Research Logistics Quarterly, 2(1-2):83–97, 1955.

J. Lee, Y. Lee, J. Kim, A. Kosiorrek, S. Choi, and Y. W. Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In International Conference on Machine Learning, pages 3744–3753. PMLR, 2019.

Q. Li, Z. Han, and X.-M. Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In Proceedings of the AAAI Conference on Artificial Intelligence, 2018.

F. Locatello, D. Weissenborn, T. Unterthiner, A. Mahendran, G. Heigold, J. Uszkoreit, A. Dosovitskiy, and T. Kipf. Object-centric learning with slot attention. In Advances in Neural Information Processing Systems, pages 11525–11538, 2020.

M. C. Mozer. A focused back-propagation algorithm for temporal pattern recognition. Complex systems, 3(4):349–381, 1989.

H. Nt and T. Maehara. Revisiting graph neural networks: All we have is low-pass filters. arXiv preprint arXiv:1905.09550, 2019.

K. Oono and T. Suzuki. Graph neural networks exponentially lose expressive power for node classification. arXiv preprint arXiv:1905.10947, 2019.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In Advances in Neural Information Processing Systems, pages 8024–8035, 2019.

B. A. Pearlmutter. Gradient calculations for dynamic recurrent neural networks: A survey. IEEE Transactions on Neural Networks, 6(5):1212–1228, 1995.

F. Pineda. Generalization of back propagation to recurrent and higher order neural networks. In Advances in Neural Information Processing Systems, pages 602–611, 1987.

F. P. Preparata and M. I. Shamos. Computational geometry: an introduction. Springer Science & Business Media, 2012.

S. Rebay. Efficient unstructured mesh generation by means of delaunay triangulation and bowyer-watson algorithm. Journal of Computational Physics, 106(1):125–138, 1993.

A. Robinson and F. Fallside. The utility driven dynamic error propagation network. University of Cambridge Department of Engineering Cambridge, MA, 1987.

H. Serviansky, N. Segol, J. Shlomi, K. Cranmer, E. Gross, H. Maron, and Y. Lipman. Set2graph: Learning graphs from sets. In Advances in Neural Information Processing Systems, 2020.

J. Shlomi, P. Battaglia, and J.-R. Vlimant. Graph neural networks in particle physics. Machine Learning: Science and Technology, 2(2):021001, 2020a.
J. Shlomi, S. Ganguly, E. Gross, K. Cranmer, Y. Lipman, H. Serviansky, H. Maron, and N. Segol. Secondary vertex finding in jets with neural networks. *arXiv preprint arXiv:2008.02831*, 2020b.

P. Y. Simard, M. B. Ottaway, and D. H. Ballard. Fixed point analysis for recurrent networks. In *Advances in Neural Information Processing Systems*, pages 149–159, 1988.

O. Vinyals, M. Fortunato, and N. Jaitly. Pointer networks. In *Advances in Neural Information Processing Systems*, 2015.

P. J. Werbos. Generalization of backpropagation with application to a recurrent gas market model. *Neural Networks*, 1(4):339–356, 1988.

R. J. Williams and J. Peng. An efficient gradient-based algorithm for on-line training of recurrent network trajectories. *Neural computation*, 2(4):490–501, 1990.

E. Winston and J. Z. Kolter. Monotone operator equilibrium networks. In *Advances in Neural Information Processing Systems*, volume 33, pages 10718–10728, 2020.

M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. R. Salakhutdinov, and A. J. Smola. Deep sets. In *Advances in Neural Information Processing Systems*, 2017.

D. W. Zhang, G. J. Burghouts, and C. G. M. Snoek. Set prediction without imposing structure as conditional density estimation. In *International Conference on Learning Representations*, 2021.

Y. Zhang, J. Hare, and A. Prügel-Bennett. Deep set prediction networks. In *Advances in Neural Information Processing Systems*, 2019.

J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun. Graph neural networks: A review of methods and applications. *AI Open*, 1:57–81, 2020.
A Experiments details

In this section we provide further details on the experimental setup and additional results.

**Particle partitioning** Both $\phi_V$ and $\phi_E$ are instantiated as a 2-layer equivariant DeepSets [Zaheer et al. [2017]]:

$$\text{Equivariant affine map} \rightarrow \text{ReLU} \rightarrow \text{Layer norm} \rightarrow \text{Equivariant affine map.} \quad (5)$$

The $\phi_I$ is instantiated as a 2-layer fully-connected neural network:

$$\text{Affine map} \rightarrow \text{ReLU} \rightarrow \text{Affine map} \rightarrow \text{Sigmoid.} \quad (6)$$

We simplify the hyperparameter search by choosing the same number $d$ for the vertices $d_V$, edges $d_E$ and all output dimensions of the affine maps in (5) and (6). In all runs dedicated to searching $d$, we set the number of total iterations $T_{\text{total}} = 3$ and backpropagate through all iterations. We start with $d=32$ and double it, until an increase yields no substantial performance gains on the validation set, resulting in $d=128$. We train all models using the Adam optimizer [Kingma and Ba [2014]] with a learning rate of 0.0003 for 400 epochs and retain the parameters corresponding to the lowest validation loss. We apply the same $d=128$ to both the Slot Attention and Set Transformer baselines. Similar to the original version [Locatello et al., 2020], we train Slot Attention with 3 iterations. Attempts with more than 3 iterations resulted in frequent divergences in the training losses. We attribute this behavior to the recurrent sinkhorn operation, that acts as a contraction map, forcing all slots to the same vector in the limit. Our model has 268162 trainable parameters, similar to 251906 for the Slot Attention baseline, but less than 517250 for Set Transformer and 461289 for Set2Graph [Serviansky et al., 2020]. The total training time is less than 12 hours on a single GTX 1080 Ti and 10 CPU cores.

For completeness, we also report the results for the rand index (RI) in Table 4.

| Model          | bottom jets | charm jets | light jets |
|---------------|-------------|------------|------------|
| Set2Graph     | 0.736±0.004 | 0.727±0.003 | 0.970±0.001 |
| Set Transformer| 0.734±0.004 | 0.734±0.004 | 0.967±0.002 |
| Slot Attention | 0.703±0.013 | 0.714±0.009 | 0.958±0.003 |
| Ours          | 0.781±0.002 | 0.751±0.001 | 0.969±0.001 |

**Convex hull finding** We apply the same hyperparameters, architectures and optimizer as in the particle partitioning experiment, except for $T_{\text{total}}$ and $N_{\text{BPTT}}$ as we mention in Section 3. This notably differs to Set2Graph, which reports an increased parameter count of 1186689 [Serviansky et al., 2020]. We train our method until we observe no improvements on the F1 validation performance for 20 epochs, with a maximum of 1000 epochs. The set-to-set baselines are trained for 4000 epochs, and we retain the parameters resulting in the highest F1 score on the validation set. The total training time is between 14 and 50 hours on a single GTX 1080 Ti and 10 CPU cores.

**Delaunay triangulation** For Delaunay triangulation we adapt the hypergraph refiner to the graph prediction case. We replace $\phi_E$ with $\phi_V$ and apply weight sharing in the incidence MLP, to reflect the symmetry of undirected graphs. We apply the same neural networks for both $\phi_V$ and $\phi_I$ as in the previous experiments, but increase the latent dimensions to $d=256$, resulting in 595201 trainable parameters. This notably differs to Set2Graph, which increases the parameter count to 5918742 [Serviansky et al., 2020], an order of magnitude larger. The total training, for the results in Table 2 is less than 9 hours on a single GTX 1080 Ti and 10 CPU cores.

**BPTT with skips** In general, truncated BPTT allows for overlaps between subsequent BPTT applications, as we illustrate in Figure 5. We compare BPTT with skips to truncated BPTT [Williams and Peng, 1990] with $T_{\text{BPTT}} = 4$ every 4 iterations, which is the setting that is most similar to ours.
with regard to training time. The BPTT results serve as a reference point to answer the question: "What if we apply BPTT more frequently, resulting in a better approximation to the true gradients?", without necessitating a grid search over all possible hyperparameter combinations for truncated BPTT.

Figure 5: Adding gradient skips to BPTT (a) Standard BPTT (b) Truncated BPTT, applying BPTT on 4 iterations every 2nd iteration (c) BPTT with skips at iterations 1, 6, 7, 8, which effectively reduces the training time, while retaining the same number of refinement steps.