

Graph Neural Networks for Massive MIMO Detection

Andrea Scotti^{1,2} Nima N. Moghadam² Dong Liu¹ Karl Gafvert² Jinliang Huang²

Abstract

In this paper, we innovately use graph neural networks (GNNs) to learn a message-passing solution for the inference task of massive multiple multiple-input multiple-output (MIMO) detection in wireless communication. We adopt a graphical model based on the Markov random field (MRF) where belief propagation (BP) yields poor results when it assumes a uniform prior over the transmitted symbols. Numerical simulations show that, under the uniform prior assumption, our GNN-based MIMO detection solution outperforms the minimum mean-squared error (MMSE) baseline detector, in contrast to BP. Furthermore, experiments demonstrate that the performance of the algorithm slightly improves by incorporating MMSE information into the prior.

1. Introduction

Massive MIMO (multiple-input and multiple-output) is a method to improve the spectral efficiency and link reliability of wireless communication systems (Goldsmith et al., 2003), by having a large number of transmitter and receiver antennas. In the fifth-generation (5G) mobile communication system, massive MIMO is a key technology to face the increasing number of mobile users and satisfy user demands. One of the challenging problems in massive MIMO is to design efficient detection algorithms for recovering the transmitted information from multiple users (Albreem et al., 2019). The optimal solution for the MIMO detection problem is the maximum likelihood (ML) detector (Albreem et al., 2019). However, ML detection is not used in practice because its complexity increases exponentially with the number of transmitters. A number of sub-optimal solutions have been proposed to balance the trade-off between performance and complexity, e.g., sphere decoding (SD) (Guo & Nilsson, 2006), zero-forcing (ZF) and minimum

mean-squared error (MMSE) detectors (Xie et al., 1990), etc. In the last decade, methods based on probabilistic graphical models (PGM) have been actively studied (Goldberger & Leshem, 2009; Goldberger & Leshem, 2010; Liu et al., 2019), where the MIMO detection problem is firstly modeled by a maximum a posteriori (MAP) inference task in a pairwise Markov random field (MRF) and then addressed approximately with belief propagation (BP) (Yedidia et al., 2003). BP is an iterative message-passing algorithm for performing exact inference on tree-structured graphical models. Its low complexity and efficiency, even for general graphs, make it very attractive for massive MIMO detection. However, due to dense connections in the MRF graph representation of the MIMO problem, BP's performance is sensitive to both prior information and the message update rules.

In this work, we innovatively use graph neural networks (GNNs) to learn a message-passing solution that addresses the inference task of MIMO detection. Specifically, our approach is built upon the GNN framework developed in (Yoon et al., 2018). Instead of propagating messages by hand-crafted update functions as in BP, (Yoon et al., 2018) uses neural networks to learn the message-passing rules and give approximate updates.

Our network is called MIMO-GNN and it can solve MIMO detection under time-varying channels and higher-order quadrature amplitude modulation (QAM), such as 16-QAM. In practice, the correlation in the channel is not known a priori. Therefore, MIMO-GNN is trained on independent and identically distributed (i.i.d.) and Gaussian distributed channels and then tested on correlated channels drawn from a different distribution (specifically, the Kronecker model (Loyka, 2001)).

Notations— We denote the transpose, the (i, j) entry and the j -th column of matrix \mathbf{A} , by \mathbf{A}^T , a_{ij} and \mathbf{a}_j , respectively. a_i stands for the i -th entry of vector \mathbf{a} . \mathbf{I}_N denotes the identity matrix of shape N .

2. Background

2.1. MIMO Detection

Wireless communication in a MIMO system requires coordination of multiple antennas at the receiver unit to detect the signals sent from wireless devices. These devices op-

¹KTH Royal Institute of Technology, Stockholm, Sweden

²Huawei Technologies Sweden AB, Stockholm, Sweden. Correspondence to: Andrea Scotti <scotti@kth.se>.

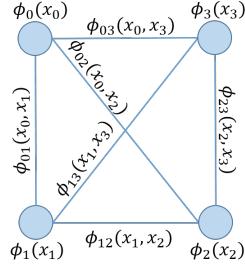


Figure 1. Fully-connected pair-wise Markov Random Fields with 4 variables.

erate as mobile transmitters within a limited coverage area commonly known as cell. Here, we consider the uplink communication in a cellular system where the base station has the role of a central coordinator. The MIMO system described above can be modeled by the real-valued linear system

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}. \quad (1)$$

The goal of MIMO detection is to infer the transmitted signal vector $\mathbf{x} \in \mathcal{A}^{N_t}$ where $\mathcal{A} \subset \mathbb{R}$ is a discrete finite alphabet ($|\mathcal{A}| = \sqrt{M}$ according to M -QAM) and N_t is the number of transmitted symbols. The channel matrix $\mathbf{H} \in \mathbb{R}^{N_r \times \mathbb{R}^{N_t}}$ and measurement vector $\mathbf{y} \in \mathbb{R}^{N_r}$ are known variables where N_r is the number of received symbols. The noise vector $\mathbf{n} \in \mathbb{R}^{N_r}$ is zero-mean Gaussian $\mathbf{n} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_{N_r})$. The real-valued system described above is derived from the actual complex-valued system where the number of receiver and transmitter antennas are $\frac{N_r}{2}$ and $\frac{N_t}{2}$ respectively. More details regarding the conversion from the complex-valued to the real-valued system is provided in (Goldberger & Leshem, 2009).

2.2. Pair-wise MRF

A MRF models the structured dependency of a set of random variables $\mathbf{x} = \{x_0, \dots, x_{N-1}\}$ by an undirected graph $\mathcal{G} = \{V, E\}$, where V and E are the set of nodes and edges respectively. Every node $i \in V$ is associated to variable x_i and it holds that $p(x_i | \mathbf{x} \setminus x_i) = p(x_i | ne(i))$, where \ denotes exclusion and $ne(i)$ is the set of neighbors of node i . In a pair-wise MRF a self potential $\phi_i(x_i)$ is assigned to node $i \in V$ and a pair potential $\phi_{ij}(x_i, x_j)$ is assigned to the edge $e \in E$ that connects node $i \in V$ to node $j \in V$. The probability distribution corresponding to a pair-wise MRF has the following form:

$$p(\mathbf{x}) = \frac{1}{Z} \prod_{i \in V} \phi_i(x_i) \prod_{(i,j) \in E} \phi_{ij}(x_i, x_j), \quad (2)$$

where Z is a normalization constant.

In order to obtain an approximation $b_l(x_l)$ of the marginal distribution for the variable x_l , we can run the iterative message-passing algorithm, BP (Yedidia et al., 2003).

2.3. MIMO as a Markov Random Field

Given the constrained linear system in (1), the corresponding posterior probability $p(\mathbf{x} | \mathbf{y})$ is factorized according to the Bayes's rule in the following way:

$$p(\mathbf{x} | \mathbf{y}) \propto p(\mathbf{y} | \mathbf{x})p(\mathbf{x}) = \exp \left\{ -\frac{1}{2\sigma^2} \|\mathbf{H}\mathbf{x} - \mathbf{y}\|^2 \right\} p(\mathbf{x}), \quad (3)$$

where $p(\mathbf{x})$ is the prior distribution for \mathbf{x} . The goal of MIMO detection is to solve the following MAP problem:

$$\hat{\mathbf{x}}_{MAP} = \arg \max_{\mathbf{x} \in \mathcal{A}^{N_t}} p(\mathbf{x} | \mathbf{y}). \quad (4)$$

The posterior probability in (3) can be factorized into a pair-wise MRF as in (2) by assignment

$$\phi_i(x_i) = e^{\frac{1}{\sigma^2} (\mathbf{y}^T \mathbf{h}_i x_i - \frac{1}{2} \mathbf{h}_i^T \mathbf{h}_i x_i^2)} p_i(x_i), \quad (5)$$

$$\phi_{ij}(x_i, x_j) = e^{-\frac{1}{\sigma^2} \mathbf{h}_i^T \mathbf{h}_j x_i x_j}, \quad (6)$$

where σ^2 is the noise variance and \mathbf{h}_i is the i -column of \mathbf{H} . By applying the BP algorithm, where the initial messages are the uniform prior probabilities over symbols, we can approximate the solution of the MAP problem in Equation (4) by solving a simplified MAP problem for each variable. Indeed, after convergence of BP, for each variable x_l we compute the belief $b_l(x_l)$ as a function of the updated messages (Yedidia et al., 2003) and hard detect the transmitted symbol \hat{x}_l with

$$\hat{x}_l = \arg \max_{x_l \in \mathcal{A}} b_l(x_l), \quad \forall l. \quad (7)$$

2.4. GNNs

GNNs from (Yoon et al., 2018) combine the advantages of deep learning and MRFs in a unique framework to capture the structure of the data into feature vectors that are updated through message-passing between nodes. In a GNN, a vector $\mathbf{u}_i \in \mathbb{R}^{S_u}$, where S_u is a positive integer, encodes the information of a variable node in a MRF (2). The values of $\{\mathbf{u}_i\}$ are iteratively updated by a recurrent neural network (RNN) with input including the value of \mathbf{u}_i at the previous iteration together with the information coming from the neighbor nodes states $\mathbf{u}_j : j \in ne(i)$ on the specified graph \mathcal{G} defined in Section 2.2.

The network is composed by three main modules: a propagation, an aggregation and a readout module. The first two modules operate at every iteration t while the readout module is involved only after the last iteration T . The propagation module outputs the updated message $\mathbf{m}_{i \rightarrow j}^t$ for each direct edge $e_{ij} \in E$

$$\mathbf{m}_{i \rightarrow j}^t = M(\mathbf{u}_i^{t-1}, \mathbf{u}_j^{t-1}, \epsilon_{ij}), \quad (8)$$

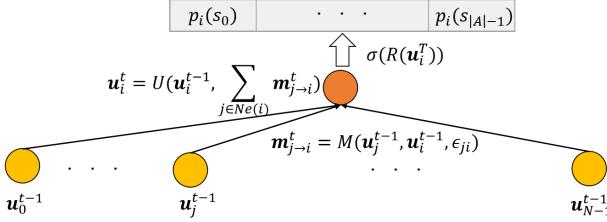


Figure 2. Message, state and output updates in GNNs.

where ϵ_{ij} is the information associated to the edge e_{ij} and M is a multiple layer perceptron (MLP) with ReLU as activation functions. Therefore, the information exchanged between two nodes at iteration t is an encoding of the concatenation of the feature vectors of the two nodes and the information along the direct edge between them. The aggregation module operates at a node level by aggregating the incoming messages $\mathbf{m}_{j \rightarrow i}^t$ at node $i \in V$, with $j \in ne(i)$, by following

$$\mathbf{u}_i^t = U(\mathbf{u}_i^{t-1}, \sum_{j \in ne(i)} \mathbf{m}_{j \rightarrow i}^t), \quad (9)$$

where U is a GRU (Cho et al., 2014).

After T iterations, the feature vectors \mathbf{u}_i are used to make inference with the readout module. If the problem that we want to solve is to compute the marginal probabilities of discrete random variables, the readout module is a MLP R of the feature vector \mathbf{u}_i followed by the softmax function $\gamma : \mathbb{R}^{|A|} \rightarrow \mathbb{R}^{|A|}$. The softmax function maps the non-normalized output \mathbf{z} of the network R to a probability distribution over predicted output symbols s_k

$$\hat{p}(x_i = s_k) = \gamma(\mathbf{z})_k = \frac{e^{z_k}}{\sum_{j=1}^{|A|} e^{z_j}}. \quad (10)$$

The parameters of M , U and R are shared across the whole graph and we learn them with supervised learning by minimizing the loss function between the true probabilities $p(x_i)$ and the predicted ones $\hat{p}(x_i)$. A good candidate for the loss function L is the cross-entropy:

$$L = - \sum_{x_i} p(x_i) \log \hat{p}(x_i), \quad (11)$$

where $\hat{p}(x_i)$ is the output of the T -layers GNN.

3. Algorithms Design

3.1. MIMO-GNN

The GNN framework presented in Section 2.4 can be used to infer the a posteriori probability $p(\mathbf{x}|\mathbf{y})$ and recover the transmitted symbols \mathbf{x} in the MIMO detection problem in (1). In this case, GNNs are built upon the MIMO MRF presented in Section 2.3. Indeed, the input of GNNs is extracted from $\phi_i(x_i)$ and $\phi_{ij}(x_i, x_j)$. The information ϵ_{ij}

along each edge e_{ij} is the feature vector $\epsilon_{ij} = [-\mathbf{h}_i^T \mathbf{h}_j, \sigma^2]$. The hidden vector \mathbf{u}_i^0 of each node i is initialized with $\mathbf{u}_i^0 = \mathbf{W}[\mathbf{y}^T \mathbf{h}_i, \mathbf{h}_i^T \mathbf{h}_i, \sigma^2]^T + \mathbf{b}$. Since we want to work with an hidden state of a given size S_u , to simplify the implementation we encode the initial vector $[\mathbf{y}^T \mathbf{h}_i, \mathbf{h}_i^T \mathbf{h}_i, \sigma^2]$ with a linear transformation given by a learnable matrix $\mathbf{W} \in \mathbb{R}^{S_u} \times \mathbb{R}^3$ and a learnable vector $\mathbf{b} \in \mathbb{R}^{S_u}$. The functions M and R are two different neural networks with two hidden layers and ReLU as activation functions. Both M and R implement dropout between hidden layers (rate of 0.1 in M and rate of 0.2 in R). The outputs of the first and the second hidden layer have sizes l and $\frac{l}{2}$ respectively. Instead, the function U is composed of a GRU network followed by a linear layer that ensures that the output size is equal to S_u . In the experiments the dimension of the GRU hidden state is l .

Since in modern wireless communication systems soft symbols are more suitable than predicted symbols without probabilistic information, the predicted value \hat{x}_l for the transmitted symbol is the expected value of x_l with probability distribution $\hat{p}(x_l)$:

$$\hat{x}_l = \mathbb{E}_{x_l} \{x_l\} = \sum_{s \in \mathcal{A}} s \hat{p}(s). \quad (12)$$

Similarly to BP for the fully connected pair-wise MRF, the complexity of MIMO-GNN is proportional to the number of edges in every iteration. However, for each edge, we need to perform a forward step in a feed-forward neural network, which increases the overall complexity.

3.2. MIMO-GNN-MMSE

In the previous section, we solve MIMO detection by assuming a uniform prior $p(\mathbf{x})$ over the unknown symbols \mathbf{x} . In this section, to improve the prior information, we incorporate the MMSE posterior as the prior $p(\mathbf{x})$ such that

$$p_l(x_l) = \frac{1}{\sqrt{2\pi c_{ll}}} \exp \left(-\frac{(z_l - x_l)^2}{2c_{ll}} \right), \quad (13)$$

where z_l is the l -th element of the MMSE estimation vector $\mathbf{z} = (\mathbf{H}^T \mathbf{H} + \sigma^2 \mathbf{I}_{N_t})^{-1} \mathbf{H}^T \mathbf{y}$ and c_{ll} is the (l, l) element of $\mathbf{C} = \sigma^2 (\mathbf{H}^T \mathbf{H} + \sigma^2 \mathbf{I}_{N_t})^{-1}$. The prior correlation coefficient ρ_{ij} between the variable x_i and x_j , $\rho_{ij} = \frac{c_{ij}^2}{c_{ii} c_{jj}}$, is added to the feature vector ϵ_{ij} .

In the implementation we reuse the same model in Section 3.1 (with the same hyperparameters) and we only modify the information ϵ_{ij} along the edges and the initial value of the hidden states \mathbf{u}_i^0 . The information along each edge e_{ij} becomes $\epsilon_{ij} = [\rho_{ij}, -\mathbf{h}_i^T \mathbf{h}_j, \sigma^2]$, and the initial hidden vector \mathbf{u}_i^0 of each node i is initialized with $\mathbf{u}_i^0 = \mathbf{W}[z_i, c_{ii}, \mathbf{y}^T \mathbf{h}_i, \mathbf{h}_i^T \mathbf{h}_i, \sigma^2]^T + \mathbf{b}$, where $\mathbf{W} \in \mathbb{R}^{S_u} \times \mathbb{R}^5$ is a learnable matrix and $\mathbf{b} \in \mathbb{R}^{S_u}$ is a learnable vector.

MIMO-GNN-MMSE exhibits a higher complexity than MIMO-GNN due to the computation of \mathbf{z} and \mathbf{C} that require the inversion of a matrix of size $N_t \times N_t$.

4. Numerical Experiments

We consider a MIMO configuration with 16 transmitter antennas ($N_t = 32$) and 32 receiver antennas ($N_r = 64$). The modulation scheme is 16-QAM.

To synthetically build the datasets (for training, validation and testing) we use three sources of randomness in each sample: signal \mathbf{x} , channel noise \mathbf{n} and channel matrix \mathbf{H} . We ensure that the transmitter power satisfies $\mathbb{E}\{\mathbf{x}^T \mathbf{x}\} = N_t$. The transmitted signal \mathbf{x} is generated randomly and uniformly over the corresponding constellation set. The channel noise standard deviation σ is derived from the definition of SNR:

$$\text{SNR} = 10 \log_{10} \frac{\mathbb{E}\{\|\mathbf{Hx}\|_2^2\}}{\mathbb{E}\{\|\mathbf{n}\|_2^2\}} = 10 \log_{10} \frac{\mathbb{E}\{\|\mathbf{Hx}\|_2^2\}}{N_r \sigma^2}.$$

MIMO-GNN and MIMO-GNN-MMSE are both trained on a pre-built dataset of size 65536 and batch size 64. The size of the (additional) validation dataset is 25% of the training dataset size. The noise standard deviation σ is fixed within each batch. Since the dataset labels must be a discrete probability distribution $p(s)$ over the constellation symbols $s \in \mathcal{A}^{N_t}$, we opt for one-hot encoded labels where $p(s) = 1$ when $s = x_l$ and 0 otherwise, where x_l is the transmitted symbol.

MIMO-GNN and MIMO-GNN-MMSE are both trained with early stopping, Adam optimizer and a learning rate of 0.0001 to minimize the loss L defined in (11). Since the correlation in the channel is not known a priori, the training is performed over channel matrices randomly sampled from the i.i.d. Gaussian channel model, where it holds that $h_{ij} \sim \mathcal{N}(0, \frac{1}{N_r})$ for each element of \mathbf{H} . After cross-validation, the hyperparameters are chosen to be $l = 128$, $S_m = S_u = 8$, $T = 10$.

MMSE, BP, MIMO-GNN and MIMO-GNN-MMSE are tested on the Kronecker channel model that controls the correlation in the MIMO channel through a correlation coefficient ρ , according to the exponential correlation model (Loyka, 2001) that structures the channel matrix \mathbf{H} as follows: $\mathbf{H} = \mathbf{R}_R^{1/2} \mathbf{K} \mathbf{R}_T^{1/2}$. Here, $k_{ij} \sim \mathcal{N}(0, \frac{1}{N_r})$ and $\mathbf{R}_R, \mathbf{R}_T$ are the spatial correlation matrices at the receiver and the transmitter side respectively.

The performances of the algorithms are tested according to the symbol error rate (SER) metric. The results are averaged over a (additional) dataset of 20000 random simulations. BP runs for 8 iterations and implements a damping factor of 0.75 on belief and messages (Som et al., 2010) to increase the performance. Moreover, the prior $p_l(x_l)$ at iteration $t+1$ is improved with the belief $b_l(x_l)$ computed at iteration t .

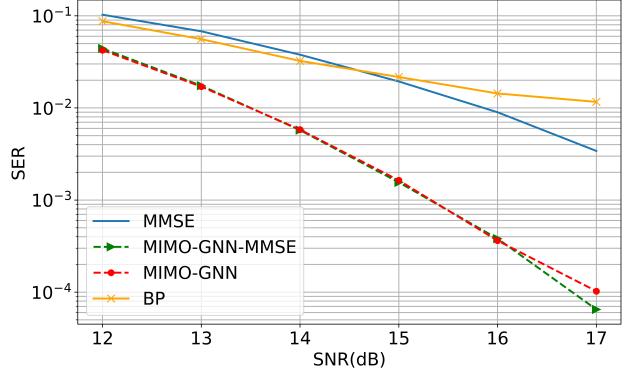


Figure 3. SER vs. SNR of different schemes for 16-QAM modulation, MIMO system with $N_t = 32$ and $N_r = 64$ and channels i.i.d. and Gaussian distributed.

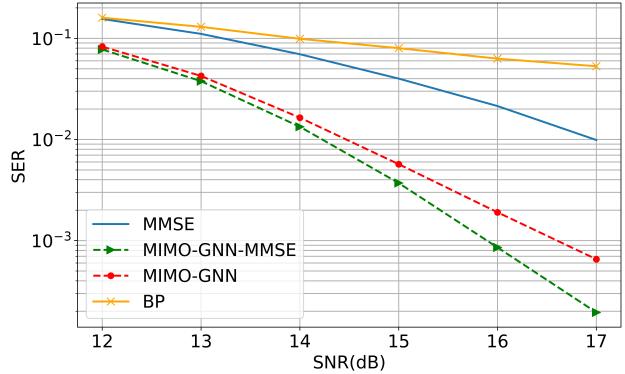


Figure 4. SER vs. SNR of different schemes for 16-QAM modulation, MIMO system with $N_t = 32$ and $N_r = 64$ and channels randomly sampled from Kronecker model with $\rho = 0.3$.

Fig. 3 shows the results for i.i.d. and Gaussian distributed channels (Kronecker model with $\rho = 0$). The performance gain of MIMO-GNN over MMSE is approximately 2.5dB at SER = 10^{-2} . Meanwhile, the improvement of MIMO-GNN-MMSE over MIMO-GNN is negligible. While, Fig. 4 shows the results for correlated channels with $\rho = 0.3$. MIMO-GNN maintains around 2dB gain over MMSE when SER is 10^{-2} . MIMO-GNN-MMSE outperforms MIMO-GNN in all the SNR range of the experiments. Integrating the MMSE prior in the model helps to increase of 0.5dB the performance gain when SER is 10^{-3} .

5. Conclusions

We have developed MIMO-GNN, a GNN-based algorithm to solve massive MIMO detection at higher-order modulation. In contrast with BP, our experiments show that the uniform prior is sufficiently informative for MIMO-GNN to significantly outperform MMSE. This performance gain, even on correlated channels, makes MIMO-GNN a promising solution for MIMO detection. Moreover, since the computation in each iteration is done independently for every edge of the graph, the complexity of our solution can be considerably reduced by a parallelization of the algorithm.

References

- Albreem, M. A., Juntti, M., and Shahabuddin, S. Massive MIMO detection techniques: A survey. *IEEE Communications Surveys Tutorials*, 21(4):3109–3132, 2019.
- Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. Learning phrase representations using RNN encoderdecoder for statistical machine translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014. doi: 10.3115/v1/d14-1179.
- Goldberger, J. and Leshem, A. A Gaussian tree approximation for integer least-squares. In Bengio, Y., Schuurmans, D., Lafferty, J. D., Williams, C. K. I., and Culotta, A. (eds.), *Advances in Neural Information Processing Systems 22*, pp. 638–645. Curran Associates, Inc., 2009
- Goldberger, J. and Leshem, A. Pseudo prior belief propagation for densely connected discrete graphs. In *2010 IEEE Information Theory Workshop on Information Theory (ITW 2010, Cairo)*, pp. 1–5, Jan 2010. doi: 10.1109/ITWKSPS.2010.5503198.
- Goldsmith, A., Jafar, S. A., Jindal, N., and Vishwanath, S. Capacity limits of MIMO channels. *IEEE Journal on Selected Areas in Communications*, 21(5):684–702, 2003.
- Guo, Z. and Nilsson, P. Algorithm and implementation of the k-best sphere decoding for MIMO detection. *IEEE Journal on selected areas in communications*, 24(3):491–503, 2006.
- Liu, D., Moghadam, N. N., Rasmussen, L. K., Huang, J., and Chatterjee, S. α belief propagation as fully factorized approximation, 2019.
- Loyka, S. L. Channel capacity of MIMO architecture using the exponential correlation matrix. *IEEE Communications Letters*, 5(9):369–371, 2001.
- Som, P., Datta, T., Chockalingam, A., and Rajan, B. S. Improved large-mimo detection based on damped belief propagation. In *2010 IEEE Information Theory Workshop on Information Theory (ITW 2010, Cairo)*, pp. 1–5, 2010.
- Xie, Z., Short, R. T., and Rushforth, C. K. A family of suboptimum detectors for coherent multiuser communications. *IEEE Journal on Selected Areas in Communications*, 8(4):683–690, 1990.
- Yedidia, J. S., Freeman, W. T., and Weiss, Y. Understanding belief propagation and its generalizations. In *IJCAI 2003*, 2003.
- Yoon, K., Liao, R., Xiong, Y., Zhang, L., Fetaya, E., Urtasun, R., Zemel, R., and Pitkow, X. Inference in probabilistic graphical models by graph neural networks, 2018.