Local Inconsistency Resolution: The Interplay between Attention and Control in Probabilistic Models

Anonymous Author
Anonymous Institution

Abstract

We present a generic algorithm for learning and approximate inference with an intuitive epistemic interpretation: iteratively focus on a subset of the model and resolve inconsistencies using the parameters under control. This framework, which we call Local Inconsistency Resolution (LIR) is built upon Probabilistic Dependency Graphs (PDGs), which provide a flexible representational foundation capable of capturing inconsistent beliefs. We show how LIR unifies and generalizes a wide variety of important algorithms in the literature, including the Expectation-Maximization (EM) algorithm, belief propagation, adversarial training, GANs, and GFlowNets. Each of these methods can be recovered as a specific instance of LIR by choosing a procedure to direct focus (attention and control). We implement this algorithm for discrete PDGs and study its properties on synthetically generated PDGs, comparing its behavior to the global optimization semantics of the full PDG.

1 Introduction

2 Preliminaries and Parametric PDGs

3 The Local Inconsistency Resolution (LIR) Algorithm

4 Unifying Classical Algorithms as Instances of LIR

4.1 The Classification Setting

4.2 The EM Algorithm and VAEs

4.3 Generative Adversarial Networks

4.4 Message Passing

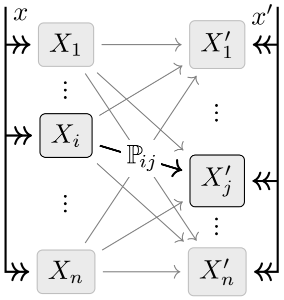
4.5 Transformer Layers

The key innovation of the transformer architecture (Vaswani et al., 2017), the basis of modern language models, is that of (scaled dot-product) (masked) *self-attention*. This notion of attention can be viewed as an instance of our framework.

Suppose that we are looking at a sequence of tokens n, whose current representations at some layer of the model are the vectors $x_1, \ldots, x_n \in \mathbb{R}^d$. Let $x'_1, \ldots x'_n$ denote the transformed representations of these tokens.

Consider a PDG with variables $\mathcal{X} = \{X_i\}_{i=1}^n \cup \{X'_i\}_{i=1}^n$. We now construct a bipartite graph, with $n^2 + 2n$ arcs; for each $(i, j) \in [n]^2$, we have an arc with cpd $\mathbb{P}_{ij} = \mathcal{N}(X'_j \mid V x_i, \sigma^2_{ij})$, and we also have arcs specifying the values of If the cpd has attention $\varphi(ij) = \exp\langle q_j, k_i \rangle$.

Preliminary work. Under review by AISTATS 2026. Do not distribute.



Proposition 1. *LIR applied to the PDG above, where $\phi_{ij} \propto \exp(x_j K Q^{\mathsf{T}} x_i/\sqrt{d})$, when controlling x', leads to $x'_j = \sum_i \Big(\exp(x_j K Q^{\mathsf{T}} x_i)/\sum_{i'} \exp(x_j K Q^{\mathsf{T}} x_{i'})\Big) V x_i$.*

4.6 Generative Flow Networks

5 Synthetic Experiments

6 Conclusion

References

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.