Meta-Learning MCMC Proposals

Tongzhou Wang∗
Facebook AI Research
tongzhou.wang.1994@gmail.com

Yi Wu
University of California, Berkeley
jxwuyi@gmail.com

David A. Moore†
Google
davmre@gmail.com

Stuart J. Russell
University of California, Berkeley
russell@cs.berkeley.edu

Abstract

Effective implementations of sampling-based probabilistic inference often require manually constructed, model-specific proposals. Inspired by recent progresses in meta-learning for training learning agents that can generalize to unseen environments, we propose a meta-learning approach to building effective and generalizable MCMC proposals. We parametrize the proposal as a neural network to provide fast approximations to block Gibbs conditionals. The learned neural proposals generalize to occurrences of common structural motifs across different models, allowing for the construction of a library of learned inference primitives that can accelerate inference on unseen models with no model-specific training required. We explore several applications including open-universe Gaussian mixture models, in which our learned proposals outperform a hand-tuned sampler, and a real-world named entity recognition task, in which our sampler yields higher final F1 scores than classical single-site Gibbs sampling.

1 Introduction

Model-based probabilistic inference is a highly successful paradigm for machine learning, with applications to tasks as diverse as movie recommendation [31], visual scene perception [17], music transcription [3], etc. People learn and plan using mental models, and indeed the entire enterprise of modern science can be viewed as constructing a sophisticated hierarchy of models of physical, mental, and social phenomena. Probabilistic programming provides a formal representation of models as sample-generating programs, promising the ability to explore a even richer range of models. Probabilistic programming language based approaches have been successfully applied to complex real-world tasks such as seismic monitoring [23], concept learning [18] and design generation [26].

However, most of these applications require manually designed proposal distributions for efficient MCMC inference. Commonly used “black-box” MCMC algorithms are often far from satisfactory when handling complex models. Hamiltonian Monte Carlo [24] takes global steps but is only applicable to continuous latent variables with differentiable likelihoods. Single-site Gibbs sampling [30, 1] can be applied to many model but suffers from slow mixing when variables are coupled in the posterior. Effective real-world inference often requires block proposals that update multiple variables together to overcome near-deterministic and long-range dependence structures. However, computing exact Gibbs proposals for large blocks quickly becomes intractable (approaching the difficulty of posterior inference), and in practice it is common to invest significant effort in hand-engineering computational tricks for a particular model.

∗Work done while the author was at University of California, Berkeley
†Work done while the author was at University of California, Berkeley

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.
(a) Two models of same structure, but with different parameters and thus different near-deterministic relations (shown in red). Naive MCMC algorithms like single-site Gibbs fail on both models due to these dependencies.

(b) To design a single proposal that works on both models in Fig. 1a, we consider this general model with variable parameters \( \alpha \) and \( \beta \) (shown in blue).

(c) Our neural proposal takes model parameters \( \alpha \) and \( \beta \) as input, and is trained to output good proposal distributions on randomly generated parameters. Therefore, it performs well for any given \( \alpha \) and \( \beta \). (For simplicity, inputs in diagram omit possible other nodes that the proposed nodes may depend on.)

(d) The neural proposal can be applied anywhere this structural pattern is present (or instantiated). The grey regions show example instantiations in this large model. (There are more.)

Figure 1: Toy example: Naive MCMC algorithms (e.g., single-site Gibbs) fail when variables are tightly coupled, requiring custom proposals even for models with similar structure but different dependency relations (Fig. 1a). Our goal is to design a single proposal that works on any model with similar local structure. We consider the general model where the dependency relations among nodes are represented by variable model parameters (Fig. 1b), and then train proposals parametrized by neural networks on models with randomly generated parameters (Fig. 1c). The trained proposal thus work on anywhere the structure is found (Fig. 1d). With proposals trained for many common motifs, we can automatically speed up inference on unseen models.

Can we build tractable MCMC proposals that are (1) effective for fast mixing and (2) ready to be reused across different models?

Recent advances in meta-learning demonstrate promising results in learning to build reinforcement learning agents that can generalize to unseen environments [7, 33, 9, 37]. The core idea of meta-learning is to generate a large number of related training environments under the same objective and then train a learning agent to succeed in all of them. Inspired by those meta-learning works, we can adopt a similar approach to build generalizable MCMC proposals.

We propose to learn approximate block-Gibbs proposals that can be reused within a given model, and even across models containing similar structural motifs (i.e., common structural patterns). Recent work recognized that a wide range of models can be represented as compositions of simple components [10], and that domain-specific models may still reuse general structural motifs such as chains, grids, rings, or trees [14]. We exploit this by training a meta-proposal to approximate block-Gibbs conditionals for models containing a given motif, with the model parameters provided as an additional input. At a high level, approach first (1) generates different instantiations of a particular motif by randomizing its model parameters, and then (2) meta-train a neural proposal “close to” the true Gibbs conditionals for all the instantiations (see Fig. 1). By learning such flexible samplers, we can improve inference not only within a specific model but even on unseen models containing similar structures, with no additional training required. In contrast to techniques that compile inference procedures specific to a given model [32, 19, 29], learning inference artifacts that generalize to novel models is valuable in allowing model builders to quickly explore a wide range of possible models.

We explore the application of our approach to a wide range of models. On grid-structured models from a UAI inference competition, our learned proposal significantly outperforms Gibbs sampling. For open-universe Gaussian mixture models, we show that a simple learned block proposal yields
performance comparable to a model-specific hand-tuned sampler, and generalizes to models more than those it was trained on. We additionally apply our method to a named entity recognition (NER) task, showing that not only do our learned block proposals mix effectively, the ability to escape local modes yields higher-quality solutions than the standard Gibbs sampling approach.

2 Related Work

There has been great interest in using learned, feedforward inference networks to generate approximate posteriors. Variational autoencoders (VAE) train an inference network jointly with the parameters of the forward model to maximize a variational lower bound [15, 5, 11]. However, the use of a parametric variational distribution means they typically have limited capacity to represent complex, potentially multimodal posteriors, such as those incorporating discrete variables or structural uncertainty.

A related line of work has developed data-driven proposals for importance samplers [25, 19, 27], training an inference network from prior samples which is then used as a proposal given observed evidence. In particular, Le et al. [19] generalize the framework to probabilistic programming, and is able to automatically generate and train a neural proposal network given an arbitrary model described in a probabilistic program. Our approach differs in that we focus on MCMC inference, allowing modular proposals for subsets of model variables that may depend on latent quantities, and exploit recurring structural motifs to generalize to new models with no additional training.

Several approaches have been proposed for adaptive block sampling, in which sets of variables exhibiting strong correlations are identified dynamically during inference, so that costly joint sampling is used only for blocks where it is likely to be beneficial [35, 34]. This is largely complementary to our current approach, which assumes the set of blocks (structural motifs) is given and attempts to learn fast approximate proposals.

Perhaps most related to our approach is recent work that trains model-specific MCMC proposals with machine learning techniques. In [29], adversarial training directly optimizes the similarity between posterior values and proposed values from a symmetric MCMC proposal. Stochastic inverses of graphical models [32] train density estimators to speed up inference. However, both approaches have limitations on applicable models and require model-specific training using global information (samples containing all variables). Our approach is simpler and more scalable, requiring only local information and generating local proposals that can be reused both within and across different models.

At a high level, our approach of learning an approximate local update scheme can be seen as related to approximate message passing [28, 12] and learning to optimize continuous objectives [2, 20].

3 Meta-Learning MCMC Proposals

We propose a meta-learning approach, using a neural network to approximate the Gibbs proposal for a recurring structural motif in graphical models, and to speed up inference on unseen models without extra tuning. Crucially our proposals do not fix the model parameters, which are instead provided as network input. After training with random model parametrizations, the same trained proposal can be reused to perform inference on novel models with parametrizations not previously observed.

Our inference networks are parametrized as mixture density networks [4], and trained to minimize the Kullback-Leibler (KL) divergence between the true posterior conditional and the proposal by sampling instantiations of the motif. The proposals are then accepted or rejected following the Metropolis-Hastings (MH) rule [1], so we maintain the correct stationary distribution even though the proposals are approximate. The following sections describe our work in greater depth.

3.1 Background

Although our approach applies to arbitrary probabilistic programs, for simplicity we focus on models represented as factor graphs. A model consists of a set of variables \( V \) as the nodes of a graph \( G = (V, E) \), along with a set of factors specifying a joint probability distribution \( p_\Psi(V) \) described by parameters \( \Psi \). In particular, this paper focuses primarily on directed models, in which the factors \( \Psi \) specify the conditional probability distributions of each variable given its parents. In undirected
models, such as the Conditional Random Fields (CRFs) in Sec. 4.3, the factors are arbitrary functions associated with cliques in the graph [16].

Given a set of observed evidence variables, inference attempts to sample from the conditional distribution on the remaining variables. In order to construct good MCMC proposals that generalize well across a variety of inference tasks, we take the advantage of recurring structural motifs in graphical models, such as grids, rings, and chains [14].

In this work, our goal is to train a neural network as an efficient expert proposal for a structural motif, with its inputs containing the local parameters, so that the trained proposal can be applied to different models. Within a motif, the variables are divided into a proposed set of variables that will be resampled, and a conditioning set corresponding to an approximate Markov blanket. The proposal network essentially maps the values of conditional variables and local parameters to a distribution over the proposed variables.

### 3.2 MCMC Proposals on Structural Motifs in Graphical Models

We associate each learned proposal with a structural motif that determines the shape of the network inputs and outputs. In general, structural motifs can be arbitrary subgraphs, but we are more interested in motifs that represent interesting conditional structure between two sets of variables, the block proposed variables \( B \) and the conditioning variables \( C \). A given motif can have multiple instantiations with a model, or even across models. As a concrete example, Fig. 2 shows two instantiations of a structural motif of six consecutive variables in a chain model. In each instantiation, we want to approximate the conditional distribution of two middle variables given neighboring four.

**Definition.** A structural motif \((B, C)\) (or motif in short) is an (abstract) graph with nodes partitioned into two sets, \( B \) and \( C \), and a parametrized joint distribution \( p(B, C) \) whose factorization is consistent with the graph structure. This specifies the functional form of the conditional \( p(B|C) \), but not the specific parameters.

A motif usually has many instantiations across many different graphical models.

**Definition.** For a graphical model \((G=(V,E), \Psi)\), an instantiation \((B_i, C_i, \Psi_i)\) of a motif \((B, C)\) includes

1. a subset of the model variables \((B_i, C_i) \subseteq V\) such that the induced subgraph on \((B_i, C_i)\) is isomorphic to the motif \((B, C)\) with the partition preserved by the isomorphism (so nodes in \( B \) are mapped to \( B_i \), and \( C \) to \( C_i \)), and
2. a subset of model parameters \( \Psi_i \subseteq \Psi \) required to specify the conditional distribution \( p_{\Psi_i}(B|C) \).

We would typically define a structural motif by first picking out a block of variables \( B \) to jointly sample, and then selecting a conditioning set \( C \). Intuitively, the natural choice for a conditioning set is the Markov blanket, \( C = MB(B) \). However, this is not a fixed requirement, and \( C \) could be either a subset or superset of it (or neither). We might deliberately choose to use some alternate conditioning set \( C \), e.g., a subset of the Markov blanket to gain a more computationally efficient proposal (with a smaller proposal network), or a superset with the idea of learning longer-range structure. More fundamentally, however, Markov blankets depend on the larger graph structure might not be consistent across instantiations of a given motif (e.g., if one instantiation has additional edges connecting \( B_i \) to other model variables not in \( C_i \)). Allowing \( C \) to represent a generic conditioning set leaves us with greater flexibility in instantiating motifs.

Formally, our goal is to learn a Gibbs-like block proposal \( q(B_i|C_i; \Psi_i) \) for all possible instantiations \((B_i, C_i, \Psi_i)\) of a structural motif that is close to the true conditional in the sense that

\[
\forall (B_i, C_i, \Psi_i), \quad \forall c_i \in \text{supp}(C_i), \quad q(B_i; c_i, \Psi_i) \approx p_{\Psi_i}(B_i|C_i = c_i). \tag{1}
\]
This provides another view of this approximation problem. If we choose the motif to have complex structures in each instantiation, the conditionals \( p_{\Psi_i}(B_i|C_i = c_i) \) can often be quite different for different instantiations, and thus difficult to approximate. Therefore, choosing what is a structural motif represents a trade-off between generality of the proposal and easiness to approximate. While our approach works for any structural motif complying with the above definition, we suggest using common structures as motifs, such as chain of certain length as in Fig. 2. In principle, recurring motifs could be automatically detected, but in this work, we focus on hand-identified common structures.

### 3.3 Parametrizing Neural Block Proposals

We choose mixture density networks (MDN) [4] as our proposal network parametrization. An MDN is a form of neural network whose outputs parametrize a mixture distribution, where in each mixture component the variables are uncorrelated.

In our case, a neural block proposal is a function \( q_\theta \) parametrized by a MDN with weights \( \theta \). The function \( q_\theta \) represents proposals for a structural motif \((B, C)\) by taking in current values of \( C_i \) and local parameters \( \Psi_i \), and outputting a distribution over \( B_i \). The goal is to optimize \( \theta \) so that \( q_\theta \) is close to the true conditional.

In the network output, mixture weights are represented explicitly. Within each mixture component, distributions of bounded discrete variables are directly represented as independent categorical probabilities, and distributions of continuous variables are represented as isotropic Gaussians with mean and variance. To avoid degenerate proposals, we threshold the variance of each Gaussian component to be at least \( 10^{-5} \).

### 3.4 Training Neural Block Proposals

**Loss function for a specific instantiation:** Given a particular motif instantiation, we use the KL divergence \( D(p_{\Psi_i}(B_i|C_i) \parallel q_\theta(B_i; C_i, \Psi_i)) \) as the measure of closeness between our proposal and the true conditional in Eq. 1. Taking into account all possible values \( c_i \in \text{supp}(C_i) \), we consider the expected divergence over \( C_i \)’s prior:

\[
\mathbb{E}_{C_i} \left[ D(p_{\Psi_i}(B_i|C_i) \parallel q_\theta(B_i; C_i, \Psi_i)) \right] = -\mathbb{E}_{B_i, C_i} \left[ \log q_\theta(B_i; C_i, \Psi_i) \right] + \text{constant}. \tag{2}
\]

The second term is independent of \( \theta \). So we define the loss function on \((B_i, C_i, \Psi_i)\) as

\[
\hat{L}(\theta; B_i, C_i, \Psi_i) = -\mathbb{E}_{B_i, C_i} \left[ \log q_\theta(B_i; C_i, \Psi_i) \right].
\]

**Meta-training over many instantiations:** To train a generalizable neural block proposal, we generate a set of random instantiations and optimize the loss function over all of them. Assuming a distribution over instantiations \( \mathcal{P} \), our goal is to minimize the overall loss

\[
L(\theta) = \mathbb{E}_{(B, C, \Psi_i) \sim \mathcal{P}} \left[ \hat{L}(\theta; B_i, C_i, \Psi_i) \right] = -\mathbb{E}_{(B, C, \Psi_i) \sim \mathcal{P}} \left[ \mathbb{E}_{B_i, C_i} \left[ \log q_\theta(B_i; C_i, \Psi_i) \right] \right], \tag{3}
\]

which is optimized with minibatch SGD in our experiments.

There are different ways to design the motif instantiation distribution \( \mathcal{P} \). One approach is to find a distribution over model parameter space, and attach the random parametrizations \( \Psi_i \) to \((B_i, C_i)\). Practically, it is also viable to find a training dataset of models that contains a large number of instantiations. Both approaches are discussed in detail and experimented in the experiment section.

**Neural block sampling:** The overall MCMC sampling procedure with meta-proposals is outlined in Algorithm 1, which supports building a library of neural block proposals trained on common motifs to speed up inference on previously unseen models.

### 4 Experiments

In this section, we evaluate our method of learning neural block proposals against single-site Gibbs sampler as well as several model-specific MCMC methods. We focus on three most common structural motifs: grids, mixtures and chains. In all experiments, we use the following guideline to design the proposal: (1) using small underlying MDNs (we pick networks with two hidden layers and elu activation [6]), and (2) choosing an appropriate distribution to generate parameters of the motif such that the generated parameters could cover the whole space as much as possible. More experiments details and an additional experiment are available in the supplementary materials.
We evaluate the performance of the trained neural block proposal on all grid BNs up to 500 nodes from UAI 2008 inference competition. In each epoch, for each latent variable, we try to identify and propose the block as in Fig. 3 with the variable located at center. If this is not possible, e.g., the variable is at boundaries or close to evidence, single-site Gibbs resampling is used instead.

Fig. 6 shows the performance of both our method and single-site Gibbs in terms of error integrated over time for all 180 models. The models are divided into three classes, grid-50, grid-75 and grid-90, according to the percentage of deterministic relations. Our neural block sampler significantly outperforms Gibbs sampler in nearly every model. We notice that the improvement is less significant as the percentage of deterministic relations increases. This is largely due to that the above proposal
structure in Fig. 3 can only easily handle dependency among the 9 proposed nodes. We expect an increased block size to yield stronger performance on models with many deterministic relations.

Furthermore, we compare our proposal against single-site Gibbs, and exact block Gibbs with identical proposal block, on grid models with different percentages of deterministic relations in Fig. 4. Single-site Gibbs performs worst on both models due to quickly getting stuck in local modes. Between the two block proposals, neural block sampling performs better in error w.r.t. time due to shorter computational time. However, because the neural block proposal is only an approximate of the true block Gibbs proposal, it is worse in terms of error w.r.t. epochs, as expected. Detailed comparisons on more models are available in the supplementary material.

Additionally, our approach can be used model-specifically by training only on instantiations within a particular model. In supplementary materials, we demonstrate that our method achieves comparable performance with a more advanced task-specific MCMC method, Inverse MCMC [32].

4.2 Gaussian Mixture Model with Unknown Number of Components

We next consider open-universe Gaussian mixture models (GMMs), in which the number of mixture components is unknown, subject to a prior. Similarly to Dirichlet process GMMs, these are typically treated with hand-designed model-specific split-merge MCMC algorithms.

Consider the following GMM. \( n \) points \( \mathbf{x} = \{ x_i \}_{i=1}^{n} \) are observed, and come uniformly randomly from one of \( M \) (unknown) active mixtures, with \( M \sim \text{Unif}\{1, \ldots, m\} \). Our task is to infer the
posterior of mixture means $\mu = \{\mu_j\}_{j=1,...,M}$, their activity indicators $v = \{v_j\}_{j=1,...,M}$, and the labels $z = \{z_i\}_{i=1,...,n}$, where $z_i$ is the mixture index $x_i$ comes from. Since $M$ is determined by $v$, in this experiment, we always calculate $M = \sum_j v_j$ instead of sampling $M$. Such GMMs have many nearly-deterministic relations, e.g., $p(v_j = 0, z_i = j) = 0$, causing vanilla single-site Gibbs failing to jump across different $M$ values. Split-merge MCMC algorithms, e.g., Restricted Gibbs split-merge (RGSM) [13] and Smart-Dumb/Dumb-Smart (SDDS) [36], use hand-designed MCMC moves to solve such issues. In our framework, it’s possible to deal with such relations with a proposal block including all of $z$, $\mu$ and $v$. However, doing so requires significant training and inference time (due to larger proposal network and larger proposal block), and the resulting proposal can not generalize to GMMs of different sizes.

In order to apply the trained proposal to differently sized GMMs, we choose the motif to propose $q_\theta$ for two arbitrary mixtures $(\mu_i, v_i)$ and $(\mu_j, v_j)$ conditioned on all other variables excluding $z$, and instead consider the model with $z$ variables collapsed. The inference task is then equivalent to first sampling $\mu$, $v$ from the collapsed model $p(\mu, v|x)$, and then $z$ from $p(z|\mu, v, x)$. We modify the algorithm such that the proposal from $q_\theta$ is accepted or rejected by the MH rule on the collapsed model. Then $z$ is resampled from $p(z|\mu, v, x)$. This approach is less sensitive to different $n$ values and performs well in variously sized GMMs. More details are available in the supplementary material.

We train with a small GMM with $m = 8$ and $n = 60$ as the motif, and apply the trained proposal on GMMs with larger $m$ and $n$ by randomly selecting 8 mixtures and 60 points for each proposal. Fig. 7 shows how the our sampler performs on GMM of various sizes, compared against split-merge Gibbs with SDDS. We notice that as model gets larger, Gibbs with SDDS mixes more slowly, while neural block sampling still mixes fairly fast and outperforms Gibbs with SDDS. Bottom right of Fig. 7 shows the trace plots of $M$ for both algorithms over multiple runs on the same observations. Gibbs with SDDS takes a long time to find a high likelihood explanation and fails to explore other possible ones efficiently. Our proposal, on the other hand, mixes quickly among the possible explanations.

4.3 Named Entity Recognition (NER) Tagging

Named entity recognition (NER) is the task of inferring named entity tags for words in natural language sentences. One way to tackle NER is to train a conditional random field (CRF) model representing the joint distribution of tags and word features [21]. In test time, we use the CRF build a chain Markov random field (MRF) containing only tags variables, and apply MCMC methods to sample the NER tags. We use a dataset of 17494 sentences from CoNLL-2003 Shared Task. The CRF model is trained with AdaGrad [8] through 10 sweeps over the training dataset.

---

Figure 8: Average F1 scores and average log likelihoods over entire test dataset. In each epoch, all variables in every test MRF is proposed roughly once for all algorithms. F1 scores are measured using states with highest likelihood seen over Markov chain traces. To better show comparison, epoch plots are cut off at 500 epochs and time plots at 12850s. Log likelihoods shown don’t include normalization constant.

Our goal is to train good neural block proposals for the chain MRFs built for test sentences. Experimenting with different chain lengths, we train three proposals, each for a motif of two, three, or four consecutive proposed tag variables and their Markov blanket. These proposals are trained on instantiations within MRFs built from the training dataset for the CRF model.

We then evaluate the learned neural block proposals on the previously unseen test dataset of 3453 sentences. Fig. 8 plots the performance of neural block sampling and single-site Gibbs w.r.t. both time and epochs on the entire test dataset. As block size grows larger, learned proposal takes more time to mix. But eventually, block proposals generally achieve better performance than single-site Gibbs in terms of both F1 scores and log likelihoods. Therefore, as shown in the figure, a mixed proposal of single-site Gibbs and neural block proposals can achieve better mixing without slowing down much. As an interesting observation, neural block sampling sometimes achieves higher F1 scores even before surpassing single-site Gibbs in log likelihood, implying that log likelihood is at best an imperfect proxy for performance on this task.

5 Conclusion

This paper proposes and explores the (to our knowledge) novel idea of meta-learning generalizable approximate block Gibbs proposals. Our meta-proposals are trained offline and can be applied directly to novel models given only a common set of structural motifs. Experiments show that the neural block sampling approach outperforms standard single-site Gibbs in both convergence speed and sample quality and achieve comparable performance against model-specialized methods. In will be an interesting system design problem to investigate, when given a library of trained block proposals, how an inference system in a probabilistic programming language can automatically detect the common structural motifs and (adaptively) apply appropriate samplers to help convergence for more general real-world applications.
References


