Monte Carlo MCMC: Efficient Inference by Sampling Factors

Sameer Singh
University of Massachusetts
140 Governors Drive
Amherst MA 01003
sameer@cs.umass.edu

Michael Wick
University of Massachusetts
140 Governors Drive
Amherst MA 01003
mwick@cs.umass.edu

Andrew McCallum
University of Massachusetts
140 Governors Drive
Amherst MA 01003
mccallum@cs.umass.edu

Abstract

Conditional random fields and other graphical models have achieved state of the art results in a variety of NLP and IE tasks including coreference and relation extraction. Increasingly, practitioners are using models with more complex structure—higher tree-width, larger fan-out, more features, and more data—rendering even approximate inference methods such as MCMC inefficient. In this paper we propose an alternative MCMC sampling scheme in which transition probabilities are approximated by sampling from the set of relevant factors. We demonstrate that our method converges more quickly than a traditional MCMC sampler for both marginal and MAP inference. In an author coreference task with over 5 million mentions, we achieve a 13 times speedup over regular MCMC inference.

1 Introduction

Conditional random fields and other graphical models are at the forefront of many natural language processing (NLP) and information extraction (IE) tasks because they provide a framework for discriminative modeling while succinctly representing dependencies among many related output variables. Previously, most applications of graphical models were limited to structures where exact inference is possible, for example linear-chain CRFs (Lafferty et al., 2001). More recently there has been a desire to include more factors, longer range dependencies and larger numbers of more sophisticated features; these include skip-chain CRFs for named entity recognition (Sutton and McCallum, 2004), higher-order models for dependency parsing (Carreras, 2007), entity-wise models for coreference (Culotta et al., 2007) and global models of relations (Yao et al., 2010). The increasing sophistication of these individual NLP components compounded with the community's desire to model these tasks jointly across cross-document considerations has resulted in graphical models for which inference is computationally prohibitive. Even popular approximate inference techniques such as loopy belief propagation and Markov chain Monte Carlo (MCMC) may be prohibitive.

MCMC algorithms such as Metropolis-Hastings (MH) are usually efficient for graphical models because the only factors needed to score a proposal are those touching the changed variables. However, if the model variables have high degree (neighbor many factors), if computation of factor scores is slow, or if each proposal modifies a substantial number of variables (e.g. to satisfy deterministic constraints, such as transitivity in coreference), then even MH can be prohibitively slow. For example, the seemingly innocuous proposal changing the type of a single entity requires scoring a linear number of factors (in the number of mentions of that entity). Often, however, the factors are somewhat *redundant*, for example, not all the mentions of the "USA" entity need to be examined to confidently conclude that it is a COUNTRY.

In this paper we propose an approximate MCMC framework that facilitates efficient inference in high-degree graphical models. In particular, we approximate the acceptance ratio in the Metropolis-Hastings algorithm by replacing the exact model scores with

a stochastic approximation. We propose two strategies for this approximation: static uniform sampling and adaptive confidence-based sampling, and demonstrate significant speedups on synthetic and real-world information extraction tasks.

MCMC is a popular method for dealing with large, dense graphical models for tasks in NLP and information extraction (Richardson and Domingos, 2006; Poon and Domingos, 2006; Poon et al., 2008; Singh et al., 2009; Wick et al., 2009). Popular probabilistic programming packages also rely on MCMC for inference and learning (Richardson and Domingos, 2006; McCallum et al., 2009), and parallel approaches to MCMC have also been recently proposed (Singh et al., 2011; Gonzalez et al., 2011). A generic method to speed up MCMC inference could have significant applicability.

2 MCMC for Graphical Models

Factor graphs represent the joint distribution over random variables by a product of factors that make the dependencies between the random variables explicit. Each (log) factor $f \in \mathcal{F}$ is a function that maps an assignment of its neighboring variables to a real number. The probability of an assignment y to the random variables, defined by the set of factors $\mathcal{F}$, is $P(y) = \frac{\exp \psi(y)}{Z}$ where $\psi(y) = \sum_{f \in \mathcal{F}} f(y)$ and $Z = \sum_y \exp \psi(y)$.

Often, computing marginal estimates of a model is computationally intractable due to the normalization constant Z, while maximum a posteriori (MAP) is prohibitive due to the search space. Markov chain Monte Carlo (MCMC) is an important tool for approximating both kinds of inference in these models. A particularly successful MCMC method for graphical model inference is Metropolis-Hastings (MH). Since sampling from the true model $P(y)$ is intractable, MH instead uses a simpler distribution $q(y'|y)$ that conditions on the current y and proposes a new state y' by modifying a few variables. This new assignment is then accepted with probability $\alpha = \min\left(1, \frac{P(y')}{P(y)} \frac{q(y|y')}{q(y'|y)}\right)$. Computing this acceptance probability is usually highly efficient because the partition function cancels, as do all the factors in the model that do not neighbor changed variables. MH can also be used for MAP inference; the acceptance probability is modified to include a temperature term: $\alpha = \min\left(1, \left(\frac{P(y')}{P(y)}\right)^{\tau}\right)$. If a cooling schedule is implemented for τ then the MH sampler for MAP inference can be seen as an instance of simulated annealing (Bertsimas and Tsitsiklis, 1993).

3 Monte Carlo MCMC

The benefit of MCMC lies in its ability to leverage the locality of the proposal. In particular, evaluation of each sample requires computing the score of all the factors that are *involved* in the change, i.e. all factors that neighbor any variable in the set that has changed. This evaluation becomes a bottleneck for tasks in which a large number of variables is involved in each proposal, or in which the model contains very high-degree variables, resulting in large number of factors, or in which computing the factor score involves an expensive computation, such as string similarity. Many of these arise naturally when performing joint inference, or representing uncertainty over the whole knowledge-base.

Instead of evaluating the log-score ψ of the model exactly, this paper proposes a Monte Carlo estimate of the log-score. In particular, if the set of factors for a given proposed change is $\mathcal{F}$, we use sampled subset of the factors $\mathcal{S} \subseteq \mathcal{F}$ as an approximation of the model score. Formally, $\psi(y) = \sum_{f \in \mathcal{F}} f(y) = |\mathcal{F}| \cdot \mathbb{E}_{\mathcal{F}}[f(y)]$ and $\psi_{\mathcal{S}}(y) = |\mathcal{F}| \cdot \mathbb{E}_{\mathcal{S}}[f(y)]$. We use $\psi_{\mathcal{S}}$ in the acceptance probability α to evaluate each sample. Since we are using a stochastic approximation to the model score, in general we expect to need more samples to converge. However, since evaluating each sample will be *much* faster ($O(|\mathcal{S}|)$ instead of $O(|\mathcal{F}|)$), we expect sampling overall to be faster. In the next sections we describe two strategies for sampling the set of factors $\mathcal{S}$.

3.1 Uniform Sampling

The most direct approach for subsampling the set of $\mathcal{F}$ is to perform uniform sampling. In particular, given a proportion parameter $0 < p \leq 1$, we select a random subset $\mathcal{S}_p \subseteq \mathcal{F}$ such that $|\mathcal{S}_p| = p \cdot |\mathcal{F}|$. Since this approach is agnostic as to the actual factors scores, $\mathbb{E}_{\mathcal{S}_p}[f] \equiv \mathbb{E}_{\mathcal{F}}[f]$. A low p leads to fast evaluation, however it may require a large number of samples due to the substantial approximation. On the other hand, although a high p will converge with fewer samples, evaluating each sample will be slow.

3.2 Confidence-Based Sampling

Selecting the best value for p is difficult, requiring analysis of the graph structure, and statistics on the distribution of the factors scores; often a difficult task for real-world applications. Further, the same value for p can result in different levels of approximation for different proposals, either unnecessarily accurate or restrictively noisy. We would prefer a strategy that adapts to the distribution of the scores.

Instead of sampling a fixed proportion, we can sample until we are confident that the current set of samples $\mathcal{S}_c$ is an accurate estimate of the true mean of $\mathcal{F}$. In particular, we maintain a running count of the sample mean $\mathbb{E}_{\mathcal{S}_c}[f]$ and variance $\sigma_{\mathcal{S}_c}$, using them to compute a confidence interval $I_{\mathcal{S}}$ around the estimate of the mean. Since the number of sampled factors $\mathcal{S}_c$ could be a substantial fraction of the set of factors $\mathcal{F}$,[1] we also incorporate *finite population control (fpc)* in our sample variance. We use the variance $\sigma^2_{\mathcal{S}_c} = \frac{1}{|\mathcal{S}_c|-1}\sum_{f\in\mathcal{S}_c}(f - \mathbb{E}_{\mathcal{S}_c}[f])^2$ to compute the interval $I_{\mathcal{S}_c} = 2z\frac{\sigma_{\mathcal{S}_c}}{\sqrt{|\mathcal{S}_c|}}\sqrt{\frac{|\mathcal{F}|-|\mathcal{S}_c|}{|\mathcal{F}|-1}}$, where $z = 1.96$, i.e. the 95% confidence interval. We iteratively sample factors without replacement from $\mathcal{F}$, until the confidence interval falls below a user specified threshold i. For proposals that contain high-variance factors, this strategy examines a large number of factors, while proposals that involve similar factors will result in fewer samples. Note that this user-specified threshold is agnostic to the graph structure and the number of factors, and instead directly reflects the distribution of the factor scores.

4 Experiments

4.1 Synthetic Entity Classification

Consider the task of classifying entities into a set of types, for example, POLITICIAN, VEHICLE, CITY, GOVERMENT-ORG, etc. For knowledge base construction, this prediction often takes place on the entity-level, as opposed to the mention-level common in traditional NLP. To evaluate the type at the entity-level, the scored factors examine features of all the entity mentions of the entity, along with the labels of all relation mentions for which it is an argument. See Yao et al. (2010) and Hoffmann et al. (2011) for examples of such models. Since a subset of the mentions can be sufficiently informative for the model, we expect our stochastic MCMC approach to work well.

We use synthetic data for such a model to evaluate the quality of marginals returned by the Gibbs sampling form of MCMC. Since the Gibbs algorithm samples each variable using a fixed assignment of its neighborhood, we represent generating a single sample as classification. We create models with a single unobserved variable (entity type) that neighbors many unary factors, each representing a single entity- or a relation-mention factor. Our synthetic models consist of random weights assigned to each of the 100 factors (generated from $\mathcal{N}(0.5, 1)$ for the true label, and $\mathcal{N}(-0.5, 1)$ for the false label).

We evaluate the previously described uniform sampling and confidence-based sampling, with several parameter values, and plot the L_1 error to the true marginals. We use the number of factors examined as a proxy for running time, as the effect of the steps in sampling are relatively negligible. The error in comparison to regular MCMC ($p = 1$) is shown in Figure 1, with standard error bars averaging over 100 models. Initially, as the sampling approach is made more stochastic (lowering p or increasing i), we see a steady improvement in the running time needed to obtain the same error tolerance. However, the amount of relative improvements slows as stochasticity is increased further, in fact for extreme values ($i = 0.05, p = 0.1$) the chains may perform worse than regular MCMC.

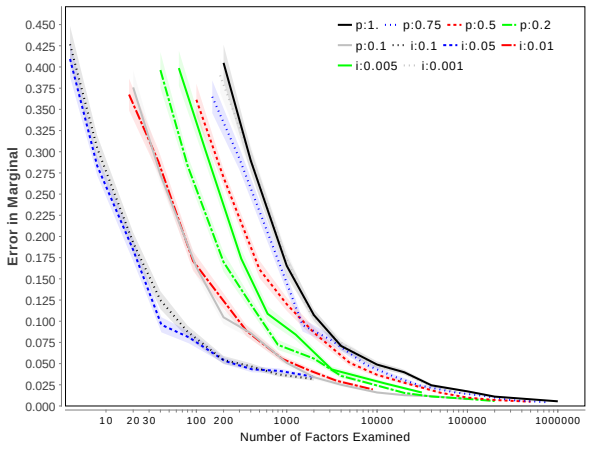

Figure 1: Synthetic Entity Classification

[1] Specifically, the fraction may be higher than 5%

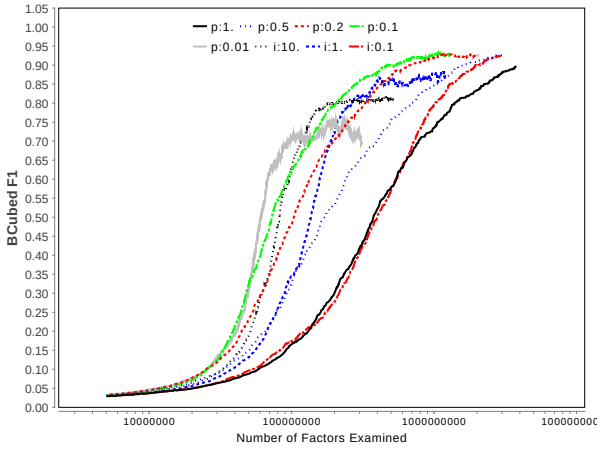

Figure 2: Entity Resolution over 5 million mentions.

Method	Factors Examined	Speedup
Baseline	1,395,330,603	1x
Uniform		
$p = 0.5$	689,254,134	2.02x
$p = 0.1$	206,157,705	6.77x
$p = 0.02$	142,689,770	9.78x
Variance		
$i = 0.1$	1,012,321,830	1.38x
$i = 1$	265,327,983	5.26x
$i = 10$	179,701,896	7.76x
$i = 100$	106,850,725	13.16x

Table 1: Speedups on DBLP to reach 80% B^3 F1

4.2 Large-Scale Author Coreference

Author coreference, the problem of clustering mentions of research paper authors into the real-world authors to which they refer, is an important step for performing meaningful bibliometric analysis. It is an instance of entity resolution, a clustering problem in which neither the identities or number of underlying entities is known. In this paper, the graphical model for entity resolution consists of observed mentions (m_i), and pairwise binary variables between pairs of mentions (y_{ij}) which represent whether the mentions are coreferent. A local factor for each y_{ij} has a high score if m_i and m_j are similar, and is instantiated only when $y_{ij} = 1$. Thus, $\psi(y) = \sum_e \sum_{m_i,m_j \in e} f(y_{ij})$. The set of possible worlds consists of all settings of the y variables such that they are consistent with transitivity, i.e. the binary variables directly represent a valid clustering over the mentions. Our proposal function selects a random mention, and moves it to a random entity, changing all the pairwise variables with mentions in its old and new entities. Thus, evaluation of such a proposal function requires scoring a number of factors linear in the size of the entities. However, the mentions are highly redundant, and observing only a subset of mentions can be sufficient.

Our dataset consists of 5 million BibTex entries from DBLP from which we extract author names, and features based on similarity between first, last names, and similarity among publication venues and co-authors. This DBLP dataset contains many large, "populous" clusters, making the evaluation of MCMC proposals computationally expensive. We also include some mentions that are labeled with their true entities, and evaluate accuracy on this subset as inference progresses. We plot *BCubed* F1, introduced by Bagga and Baldwin (1998), versus the number of factors examined (Figure 2). We also show accuracy in Table 1. We observe consistent speed improvements as stochasticity is increased. Our proposed method achieves substantial saving on this task, with a 13.16x speedup using the confidence-based sampler and 9.78x speedup using the uniform sampler. Our results also show that using extremely high confidence intervals and low sampling proportion can result in convergence to a low accuracy.

5 Conclusions

Motivated by the need for an efficient inference technique that can scale to large, densely-factored models, this paper considers a simple extension to the Markov chain Monto Carlo algorithm. By observing that many graphical models contain substantial redundancy among the factors, we propose a *stochastic* evaluation of proposals that subsamples the factors to be scored. Using two proposed sampling strategies, we demonstrate improved convergence for marginal inference on synthetic data. Further, we evaluate our approach on a large-scale, real-world entity resolution dataset, obtaining a 13x speedup on a dataset containing 5 million mentions.

Acknowledgements

We would like to thank the anonymous reviewers and Brian Martin for their valuable feedback. This work was supported in part by the Center for Intelligent Information Retrieval, in part by ARFL under prime contract number is FA8650-10-C-7059, and the University of Massachusetts gratefully acknowledges the support of Defense Advanced Research Projects Agency (DARPA) Machine Reading Program under Air Force Research Laboratory (AFRL) prime contract no. FA8750-09-C-0181. The U.S. Government is authorized to reproduce and distribute reprint for Governmental purposes notwithstanding any copyright annotation thereon. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

References

[Bagga and Baldwin1998] Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *International Conference on Language Resources and Evaluation (LREC) Workshop on Linguistics Coreference*, pages 563–566.

[Bertsimas and Tsitsiklis1993] D. Bertsimas and J. Tsitsiklis. 1993. Simulated annealing. *Statistical Science*, pages 10–15.

[Carreras2007] Xavier Carreras. 2007. Experiments with a higher-order projective dependency parser. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pages 957–961.

[Culotta et al.2007] Aron Culotta, Michael Wick, and Andrew McCallum. 2007. First-order probabilistic models for coreference resolution. In *North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT)*.

[Gonzalez et al.2011] Joseph Gonzalez, Yucheng Low, Arthur Gretton, and Carlos Guestrin. 2011. Parallel gibbs sampling: From colored fields to thin junction trees. In *Artificial Intelligence and Statistics (AISTATS)*, Ft. Lauderdale, FL, May.

[Hoffmann et al.2011] Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S. Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 541–550, Portland, Oregon, USA, June. Association for Computational Linguistics.

[Lafferty et al.2001] John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. 18th International Conf. on Machine Learning*, pages 282–289. Morgan Kaufmann, San Francisco, CA.

[McCallum et al.2009] Andrew McCallum, Karl Schultz, and Sameer Singh. 2009. FACTORIE: Probabilistic programming via imperatively defined factor graphs. In *Neural Information Processing Systems (NIPS)*.

[Poon and Domingos2006] Hoifung Poon and Pedro Domingos. 2006. Sound and efficient inference with probabilistic and deterministic dependencies. In *AAAI Conference on Artificial Intelligence*.

[Poon et al.2008] Hoifung Poon, Pedro Domingos, and Marc Sumner. 2008. A general method for reducing the complexity of relational inference and its application to MCMC. In *AAAI Conference on Artificial Intelligence*.

[Richardson and Domingos2006] Matthew Richardson and Pedro Domingos. 2006. Markov logic networks. *Machine Learning*, 62(1-2):107–136.

[Singh et al.2009] Sameer Singh, Karl Schultz, and Andrew McCallum. 2009. Bi-directional joint inference for entity resolution and segmentation using imperatively-defined factor graphs. In *Machine Learning and Knowledge Discovery in Databases (Lecture Notes in Computer Science) and European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, pages 414–429.

[Singh et al.2011] Sameer Singh, Amarnag Subramanya, Fernando Pereira, and Andrew McCallum. 2011. Large-scale cross-document coreference using distributed inference and hierarchical models. In *Association for Computational Linguistics: Human Language Technologies (ACL HLT)*.

[Sutton and McCallum2004] Charles Sutton and Andrew McCallum. 2004. Collective segmentation and labeling of distant entities in information extraction. Technical Report TR # 04-49, University of Massachusetts, July.

[Wick et al.2009] Michael Wick, Aron Culotta, Khashayar Rohanimanesh, and Andrew McCallum. 2009. An entity-based model for coreference resolution. In *SIAM International Conference on Data Mining (SDM)*.

[Yao et al.2010] Limin Yao, Sebastian Riedel, and Andrew McCallum. 2010. Collective cross-document relation extraction without labelled data. In *Empirical Methods in Natural Language Processing (EMNLP)*.