Supplementary Material for Approximate Inference by Compilation to Arithmetic Circuits

Daniel Lowd
Department of Computer and Information Science
University of Oregon
Eugene, OR 97403-1202
lowd@cs.uoregon.edu

Pedro Domingos
Department of Computer Science and Engineering
University of Washington
Seattle, WA 98195-2350
pedrod@cs.washington.edu

1 Proof of Theorem 1

Theorem 1. *For discrete probability distributions P and Q, and evidence x_{ev},*

$$KL(P(.|x_{ev}) \,\|\, Q(.|x_{ev})) \leq \frac{1}{P(x_{ev})} KL(P \,\|\, Q)$$

Proof. We will use X to refer to the set of variables whose values are set by the evidence, x_{ev}, and Y to refer to the set of all other variables.

$$\begin{aligned}
\mathrm{KL}(P(Y,X) \,\|\, Q(Y,X)) &= \sum_{y,x} P(y,x) \log \frac{P(y,x)}{Q(y,x)} && (1) \\
&= \sum_x P(x) \sum_y P(y|x) \left(\log \frac{P(y|x)}{Q(y|x)} + \log \frac{P(x)}{Q(x)} \right) && (2) \\
&= \sum_x P(x) \sum_y P(y|x) \log \frac{P(y|x)}{Q(y|x)} + \sum_x P(x) \log \frac{P(x)}{Q(x)} \sum_y P(y|x) && (3) \\
&= \sum_x P(x)\, \mathrm{KL}(P(Y|x) \,\|\, Q(Y|x)) + \mathrm{KL}(P(X) \,\|\, Q(X)) && (4) \\
&\geq P(x_{ev})\, \mathrm{KL}(P(Y|x_{ev}) \,\|\, Q(Y|x_{ev})) && (5)
\end{aligned}$$

The inequality follows from the fact that the KL divergence is non-negative. Dividing both sides by $P(x_{ev})$, we can conclude:

$$\mathrm{KL}(P(Y|x_{ev}) \,\|\, Q(Y|x_{ev})) \leq \frac{1}{P(x_{ev})} \mathrm{KL}(P(Y,X) \,\|\, Q(Y,X)) \tag{6}$$

□

Table 1: Summary statistics of the datasets and BNs used in our experiments.

Dataset	Atts.	Examples	Density	Params
KDD Cup	65	200.0k	0.008	2267
Plants	69	17.4k	0.180	2243
Audio	100	15.0k	0.198	2475
Jester	100	10.0k	0.609	2239
Netflix	100	15.0k	0.541	2532
MSWeb	294	32.7k	0.010	1710
Book	500	8.7k	0.016	2240
EachMovie	500	5.5k	0.059	4830

2 Datasets

Summary statistics of the datasets and the Bayesian networks generated from them are in Table 1. "Density" is the average fraction of true attributes in the training data; "Params" is the number of independent parameters in the learned BN. Plants consists of location data (present or not present in each state or territory) for over 22,000 plants. Anonymous MSWeb is visit data for 294 areas (Vroots) of the Microsoft Web site, collected during one week in February 1998. Both can be found in the UCI machine learning repository [1]. KDD Cup 2000 is a clickstream prediction dataset [2], which consists of Web session data taken from an online retailer. Using the subset of Hulten and Domingos [3], each example consists of 65 Boolean variables, corresponding to whether or not a particular session visited a web page matching a certain category.

EachMovie, Netflix, Audio, Book, and Jester are collaborative filtering datasets. EachMovie[1] and the Netflix consist of movie ratings; Audio[2] consists of music preferences from Audioscrobbler; Book [4] consists of book ratings from Book Crossing; and Jester [5] consists of joke ratings for 100 jokes. For the first four datasets, we reduced the variables to "rated" or "not rated" for the most popular items (or "listened to" or "not listened to" for Audio). We also reduced the number of examples by random sampling. For Book, we only selected users who had rated a certain number of books. For Jester, we selected users who had rated all 100 jokes and reduced their preferences to "like" and "dislike" by thresholding the real-valued preference ratings at zero.

Plants, Netflix, Audio, and Book datasets were processed and provided by Davis and Domingos [6].

References

[1] C. Blake and C. J. Merz. UCI repository of machine learning databases. Machine-readable data repository, Department of Information and Computer Science, University of California at Irvine, Irvine, CA, 2000. http://www.ics.uci.edu/∼mlearn/MLRepository.html.

[2] R. Kohavi, C. Brodley, B. Frasca, L. Mason, and Z. Zheng. KDD-Cup 2000 organizers' report: Peeling the onion. *SIGKDD Explorations*, 2(2):86–98, 2000.

[3] G. Hulten and P. Domingos. Mining complex models from arbitrarily large databases in constant time. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 525–531, Edmonton, Canada, 2002. ACM Press.

[4] C. Ziegler, S. McNee, J. Konstan, and G. Lausen. Improving recommendation lists through topic diversification. In *Proceedings of the 14th International World Wide Web Conference*, pages 22–32, 2005.

[5] K. Goldberg, T. Roeder, D. Gupta, and C. Perkins. Eigentaste: A constant time collaborative filtering algorithm. *Information Retrieval*, 4(2):133–151, 2001.

[6] J. Davis and P. Domingos. Bottom-up learning of Markov network structure. In *Proceedings of the Twenty-Seventh International Conference on Machine Learning*, Haifa, Israel, 2010. ACM Press.

[1] Provided by Compaq at http://research.compaq.com/SRC/eachmovie/; no longer available for download, as of October 2004.

[2] http://www-etud.iro.umontreal.ca/∼bergstrj/audio-scrobbler_data.html

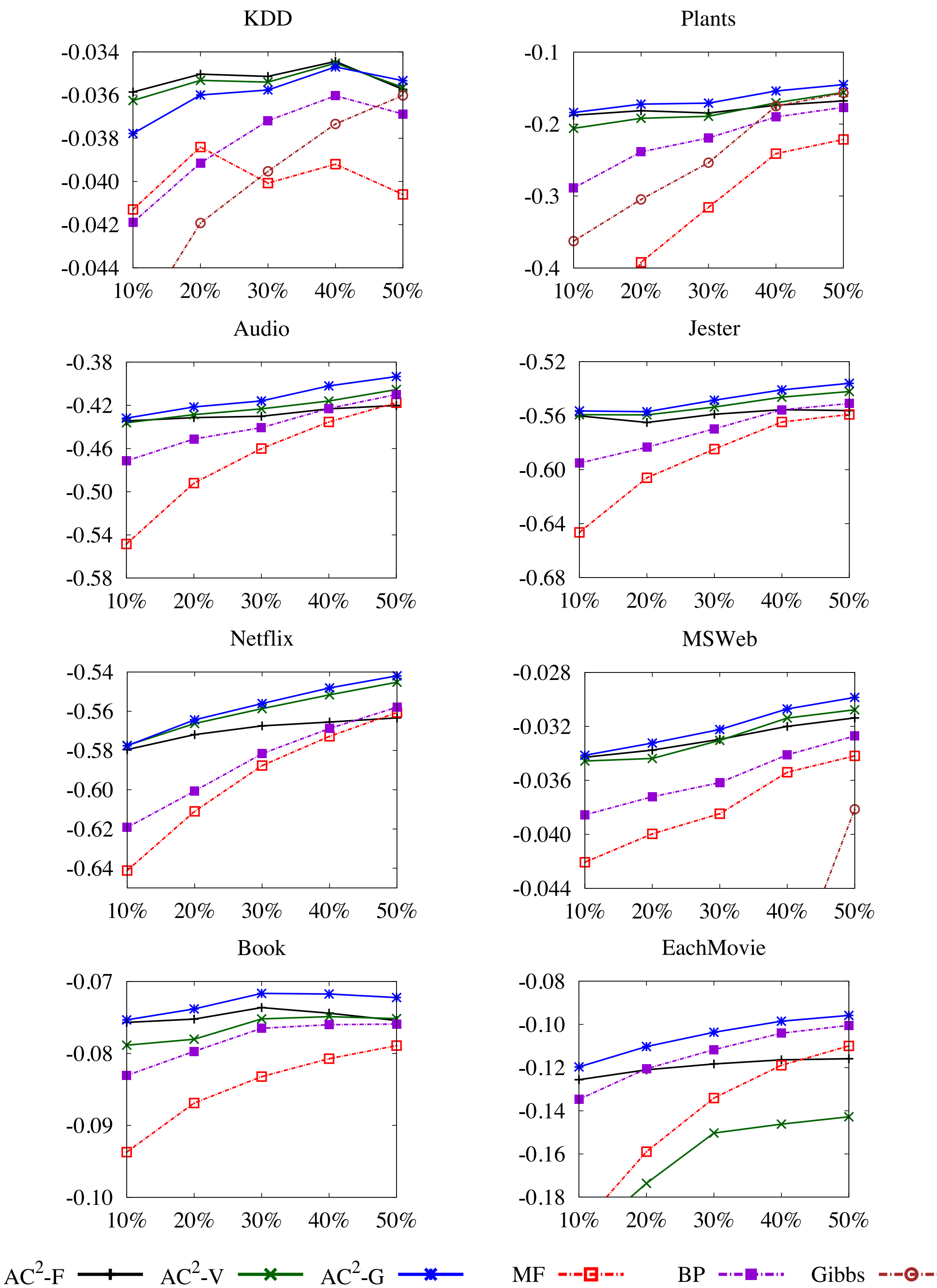

Figure 1: Average conditional log likelihood of the query variables (y axis), normalized by dividing by the number of query variables (x axis). Higher is better. Gibbs often performs too badly to appear in the frame.