FACTORIZATION, INFERENCE AND PARAMETER LEARNING IN DISCRETE AMP CHAIN GRAPHS: ADDENDUM

JOSE M. PEÑA
ADIT, IDA, LINKÖPING UNIVERSITY, SE-58183 LINKÖPING, SWEDEN
JOSE.M.PENA@LIU.SE

This note extends the original manuscript with new results, and corrects some errors.

1. Factorization

A probability distribution p is Markovian wrt an AMP CG G iff the following three properties hold for all $C \in Cc(G)$ (Andersson et al., 2001, Theorem 2):

- C1: $C \perp_p Nd_G(C) \setminus Cc_G(Pa_G(C)) | Cc_G(Pa_G(C))$.
- C2: $p(C|Cc_G(Pa_G(C)))$ is Markovian wrt G_C.
- C3*: For all $D \subseteq C$, $D \perp_p Cc_G(Pa_G(C)) \setminus Pa_G(D) | Pa_G(D)$.

Lemma 1. *C1, C2 and C3* hold iff the following two properties hold:*

- *C1*: For all $D \subseteq C$, $D \perp_p Nd_G(D) \setminus Pa_G(D) | Pa_G(D)$.*
- *C2*: $p(C|Pa_G(C))$ is Markovian wrt G_C.*

Proof. First, C1* implies C3* by decomposition. Second, C1* implies C1 by taking $D = C$ and applying weak union. Third, C1 and the fact that $Nd_G(D) = Nd_G(C)$ imply $D \perp_p Nd_G(D) \setminus Cc_G(Pa_G(C)) | Cc_G(Pa_G(C))$ by symmetry and decomposition, which together with C3* imply C1* by contraction. Finally, C2 and C2* are equivalent because $p(C|Pa_G(C)) = p(C|Cc_G(Pa_G(C)))$ by C1* and decomposition. □

Given $C \in Cc(G)$ and $D \subseteq C$, we define the marginal graph G_C^D as the undirected graph over D st $X - Y$ is in G_C^D iff $X - Y$ is in G_C or $X - V_1 - \ldots - V_n - Y$ is G_C with $V_1, \ldots, V_n \notin D$.

Lemma 2. *Assume that p is strictly positive and C1* holds. Then, C2* holds iff*

$$p(D|Pa_G(C)) = \prod_{K \in Cs(G_C^D)} \psi_D(K, Pa_G(K)) \tag{1}$$

for all $D \subseteq C$.

Proof. To prove the if part, it suffices to take $D = C$ and note that $G_C^D = G_C$. Then, C2* holds (Lauritzen, 1996, Proposition 3.8). To prove the only if part, we adapt the proof of Theorem 3.9 by Lauritzen (1996) to prove that $p(D|Pa_G(D))$ factorizes as indicated in Equation 1. This implies the desired result by C1* and decomposition. Specifically, choose arbitrary but fixed states d^* and $pa_G(D)^*$ of D and $Pa_G(D)$. Given $B \subseteq D$, let $\overline{b}^*$ and $\overline{pa_G(B)}^*$ denote the values of $D \setminus B$ and $Pa_G(D) \setminus Pa_G(B)$ consistent with d^* and $pa_G(D)^*$. For all $B \subseteq D$, let

$$H_D(b, pa_G(B)) = \log p(b, \overline{b}^* | pa_G(B), \overline{pa_G(B)}^*).$$

Note that using the logarithm is warranted by the assumption of p being strictly positive. For all $K \subseteq D$, let

$$\phi_D(k, pa_G(K)) = \sum_{B \subseteq K} (-1)^{|K \setminus B|} H_D(b, pa_G(B)) \tag{2}$$

where b is consistent with k. Now, we can apply the Möbius inversion (Lauritzen, 1996, Lemma A.2) to obtain

$$\log p(d|pa_G(D)) = H_D(d, pa_G(D)) = \sum_{K \subseteq D} \phi_D(k, pa_G(K))$$

Date: AMPCGsParameterLearning4Addendum4.tex, 14:38, 23/09/15.

where k is consistent with d. Then, it only remains to prove that $\phi_D(k, pa_G(K))$ is zero whenever $K \notin Cs(G_C^D)$. Consider two nodes S and T of K that are not adjacent in G_C^D. Then

$$\phi_D(k, pa_G(K)) = \sum_{B \subseteq K \setminus ST} (-1)^{|(K \setminus ST) \setminus B|} \big[H_D(b, pa_G(B)) - H_D(bs, pa_G(BS)) \\ - H_D(bt, pa_G(BT)) + H_D(bst, pa_G(BST)) \big] \quad (3)$$

where b, bs, bt and bst are consistent with k. Note that $S \perp_p T | (D \setminus ST) Pa_G(C)$ by C2*, and $S \perp_p Pa_G(C) \setminus Pa_G(D) | (D \setminus ST) Pa_G(D)$ by C1*, symmetry, decomposition and weak union. Then, $S \perp_p T | (D \setminus ST) Pa_G(D)$ by contraction and decomposition. This together with Equation 3.7 by Lauritzen (1996) imply that

$$H_D(bst, pa_G(BST)) - H_D(bs, pa_G(BS)) = \log \frac{p(bst, \overline{bst}^* | pa_G(BST), \overline{pa_G(BST)}^*)}{p(bs, \overline{bs}^* | pa_G(BS), \overline{pa_G(BS)}^*)}$$

$$= \log \frac{p(s|b, \overline{bst}^*, pa_G(BST), \overline{pa_G(BST)}^*) p(bt, \overline{bst}^* | pa_G(BST), \overline{pa_G(BST)}^*)}{p(s|b, \overline{bst}^*, pa_G(BS), \overline{pa_G(BS)}^*) p(b, \overline{bs}^* | pa_G(BS), \overline{pa_G(BS)}^*)}.$$

Moreover, note that $S \perp_p Pa_G(T) \setminus Pa_G(D \setminus T) | (D \setminus ST) Pa_G(D \setminus T)$ by C1*, symmetry, decomposition and weak union. This implies that

$$H_D(bst, pa_G(BST)) - H_D(bs, pa_G(BS))$$

$$= \log \frac{p(s|b, \overline{bst}^*, pa_G(BS), \overline{pa_G(BST)}^*) p(bt, \overline{bst}^* | pa_G(BST), \overline{pa_G(BST)}^*)}{p(s|b, \overline{bst}^*, pa_G(BS), \overline{pa_G(BST)}^*) p(b, \overline{bs}^* | pa_G(BS), \overline{pa_G(BS)}^*)}$$

$$= \log \frac{p(s^*|b, \overline{bst}^*, pa_G(BT), \overline{pa_G(BT)}^*) p(bt, \overline{bst}^* | pa_G(BST), \overline{pa_G(BST)}^*)}{p(s^*|b, \overline{bst}^*, pa_G(BT), \overline{pa_G(BT)}^*) p(b, \overline{bs}^* | pa_G(BS), \overline{pa_G(BS)}^*)}.$$

Moreover, $S \perp_p Pa_G(T) \setminus Pa_G(D \setminus T) | (D \setminus ST) Pa_G(D \setminus T)$ also implies that

$$H_D(bst, pa_G(BST)) - H_D(bs, pa_G(BS))$$

$$= \log \frac{p(s^*|b, \overline{bst}^*, pa_G(BT), \overline{pa_G(BT)}^*) p(bt, \overline{bst}^* | pa_G(BST), \overline{pa_G(BST)}^*)}{p(s^*|b, \overline{bst}^*, pa_G(B), \overline{pa_G(B)}^*) p(b, \overline{bs}^* | pa_G(BS), \overline{pa_G(BS)}^*)}.$$

Finally, note that $D \setminus S \perp_p Pa_G(S) \setminus Pa_G(D \setminus S) | Pa_G(D \setminus S)$ by C1* and decomposition. This implies that

$$H_D(bst, pa_G(BST)) - H_D(bs, pa_G(BS))$$

$$= \log \frac{p(s^*|b, \overline{bst}^*, pa_G(BT), \overline{pa_G(BT)}^*) p(bt, \overline{bst}^* | pa_G(BT), \overline{pa_G(BT)}^*)}{p(s^*|b, \overline{bst}^*, pa_G(B), \overline{pa_G(B)}^*) p(b, \overline{bs}^* | pa_G(B), \overline{pa_G(B)}^*)}$$

$$= \log \frac{p(bt, \overline{bt}^* | pa_G(BT), \overline{pa_G(BT)}^*)}{p(b, \overline{b}^* | pa_G(B), \overline{pa_G(B)}^*)} = H_D(bt, pa_G(BT)) - H_D(b, pa_G(B)).$$

Thus, all the terms in the square brackets in Equation 3 add to zero, which implies that the entire sum is zero. □

It is customary to think of the factors $\psi_D(K, Pa_G(K))$ in Equation 1 as arbitrary non-negative functions, whose product needs to be normalized to result in a probability distribution. Note however that Equation 1 does not include any normalization constant. The reason is that the so called canonical parameterization in Equation 2 permits us to write any probability distribution as a product of factors that does not need subsequent normalization. One might think that this must be an advantage for parameter estimation and inference. However, the truth is that the cost of computing the normalization constant has been replaced by the cost of having to manipulate a larger number of factors in Equation 1. To see it, note that the size of $Cs(G_C^D)$ is exponential in the size of the largest clique in G_C^D.

A necessary and sufficient factorization follows.

Theorem 1. *Let p be a strictly positive probability distribution. Then, p is Markovian wrt an AMP CG G iff*

$$p(V) = \prod_{C \in Cc(G)} p(C|Pa_G(C)) \tag{4}$$

with

$$p(D|Pa_G(C)) = \prod_{K \in Cs(G_C^D)} \psi_D(K, Pa_G(K)) \tag{5}$$

for all $D \subseteq C$.

Proof. The only if part holds because C1* and decomposition imply Equation 4, and Lemma 2 implies Equation 5. To prove the if part, we prove that p satisfies C1* and C2*. Note that $Nd_G(C) = Nd_G(D)$. This together with Equations 4 and 5 imply that

$$p(D, Nd_G(D)) = p(D, Nd_G(C)) = \left(\prod_{U \in Cc(G): U \subseteq Nd_G(C)} p(U|Pa_G(U)) \right) p(D|Pa_G(C))$$

$$= g(Nd_G(D)) h(D, Pa_G(D))$$

and thus C1* holds (Lauritzen, 1996, Equation 3.6). Finally, C2* holds by Equation 5 and Lemma 2. □

A more convenient necessary and sufficient factorization follows.

Theorem 2. *Let p be a strictly positive probability distribution. Then, p is Markovian wrt an AMP CG G iff*

$$p(V) = \prod_{C \in Cc(G)} p(C|Pa_G(C)) \tag{6}$$

with

$$p(C|Pa_G(C)) = \prod_{K \in Cs(G_C)} \psi_C(K, Pa_G(K)) \tag{7}$$

and

$$p(D|Pa_G(C)) = p(D|Pa_G(D)) \tag{8}$$

for all $D \subseteq C$.

Proof. The only if part holds because C1* and decomposition imply Equations 6 and 8, and Lemma 2 implies Equation 7. To prove the if part, we prove that p satisfies C1* and C2*. Note that $Nd_G(C) = Nd_G(D)$. This together with Equations 6 and 8 imply that

$$p(D, Nd_G(D)) = p(D, Nd_G(C)) = \left(\prod_{U \in Cc(G): U \subseteq Nd_G(C)} p(U|Pa_G(U)) \right) p(D|Pa_G(C))$$

$$= \left(\prod_{U \in Cc(G): U \subseteq Nd_G(C)} p(U|Pa_G(U)) \right) p(D|Pa_G(D)) = g(Nd_G(D)) h(D, Pa_G(D))$$

and thus C1* holds (Lauritzen, 1996, Equation 3.6). Finally, C2* holds by Equation 7 (Lauritzen, 1996, Proposition 3.8). □

A necessary factorization that is more convenient for inference and parameter learning follows.

Corollary 1. *Let p be a strictly positive probability distribution. If p is Markovian wrt an AMP CG G, then*

$$p(V) = \prod_{C \in Cc(G)} p(C|Pa_G(C)) \tag{9}$$

with

$$p(C|Pa_G(C)) = \prod_{K \in Cs(G_C)} \psi_C(K, Pa_G(K)). \tag{10}$$

2. Parameter Learning

Given some data, we can efficiently obtain the maximum likelihood estimates of the factors in Equation 10 by adapting the iterative proportional fitting procedure (IPFP) for MRFs (Murphy, 2012, Section 19.5.7) as follows:

1 For each $C \in Cc(G)$
2 Set $p^0(C|Pa_G(C))$ to the uniform distribution
3 Compute $\phi_C(K, Pa_G(K))$ for all $K \in Cs(G_C)$ as shown in Equation 2
4 Set $\psi_C(K, Pa_G(K)) = \exp \phi_C(K, Pa_G(K))$ for all $K \in Cs(G_C)$
5 Repeat until convergence
6 Set $\psi_C(K, Pa_G(K)) = \psi_C(K, Pa_G(K)) \frac{p_e(K|Pa_G(K))}{p(K|Pa_G(K))}$ for all $K \in Cs(G_C)$

where p_e is the empirical probability distribution over V obtained from the given data, and p is the probability distribution over V due to the current estimates. Note that computing $p(K|Pa_G(K))$ requires inference. The multiplication and division in line 6 are elementwise. Existing gradient ascend methods for MRFs can be adapted similarly.

We justify the algorithm above by adapting some existing results for MRFs. We temporally drop the assumption that the product of factors in Equation 10 is normalized, and replace it with

$$p(C|Pa_G(C)) = \frac{1}{Z_C(Pa_G(C))} \prod_{K \in Cs(G_C)} \psi_C(K, Pa_G(K)) \tag{11}$$

where

$$Z_C(Pa_G(C)) = \sum_c \prod_{K \in Cs(G_C)} \psi_C(k, Pa_G(K))$$

where k is consistent with c. Let ψ denote all the factors due to Equations 9 and 11. Then, the log-likelihood function is

$$l(\psi) = \sum_{C \in Cc(G)} \left(\sum_{K \in Cs(G_C)} \sum_k \sum_{pa_G(K)} n(k, pa_G(K)) \log \psi_C(k, pa_G(K)) - n(pa_G(C)) \log Z_C(pa_G(C)) \right)$$

where $n(k, pa_G(K))$ is the number of instances in the data where K and $Pa_G(K)$ take values k and $pa_G(K)$ simultaneously. Similarly for $n(pa_G(C))$. Dividing both sides by the number of instances in the data, n, we have that

$$l(\psi)/n = \sum_{C \in Cc(G)} \left(\sum_{K \in Cs(G_C)} \sum_k \sum_{pa_G(K)} p_e(k, pa_G(K)) \log \psi_C(k, pa_G(K)) - p_e(pa_G(C)) \log Z_C(pa_G(C)) \right).$$

Let $U \in Cc(G)$ and $Q \in Cs(G_U)$. The gradient of $l(\psi)/n$ wrt $\psi_U(q, pa_G(Q))$ is

$$\frac{\partial l(\psi)/n}{\partial \psi_U(q, pa_G(Q))} = \frac{p_e(q, pa_G(Q))}{\psi_U(q, pa_G(Q))} - \frac{p_e(pa_G(U))}{Z_U(pa_G(U))} \frac{\partial Z_U(pa_G(U))}{\partial \psi_U(q, pa_G(Q))}.$$

Let $W = U \setminus Q$. Then

$$\frac{\partial Z_U(pa_G(U))}{\partial \psi_U(q, pa_G(Q))} = \sum_w \prod_{K \in Cs(G_U) \setminus Q} \psi_U(k, \overline{k}, pa_G(K))$$

$$= \frac{Z_U(pa_G(U))}{\psi_U(q, pa_G(Q))} \sum_w \prod_{K \in Cs(G_U) \setminus Q} \psi_U(k, \overline{k}, pa_G(K)) \frac{\psi_U(q, pa_G(Q))}{Z_U(pa_G(U))} = \frac{Z_U(pa_G(U))}{\psi_U(q, pa_G(Q))} p(q|pa_G(U))$$

where $\overline{k}$ denotes the elements of q corresponding to the elements of $K \cap Q$, and the last equality follows from Equation 11. Note also that $p(q|pa_G(Q)) = p(q|pa_G(U))$ by C1* and decomposition. Putting together the results above, we have that

$$\frac{\partial l(\psi)/n}{\partial \psi_U(q, pa_G(Q))} = \frac{p_e(q, pa_G(Q))}{\psi_U(q, pa_G(Q))} - \frac{p_e(pa_G(Q)) p(q|pa_G(Q))}{\psi_U(q, pa_G(Q))}.$$

Since the maximum likelihood estimates are obtained when the gradient is 0 for all the entries of all the factors, we have that the maximum likelihood estimates are obtained when

$$\psi_C(k, pa_G(K)) = \psi_C(k, pa_G(K)) \frac{p_e(k|pa_G(K))}{p(k|pa_G(K))}$$

for all $C \in Cc(G)$ and $K \in Cs(G_C)$. This justifies the updating step in line 6 of the IPFP.

Let the factor updated in the current iteration have the superscript $t+1$, whereas the rest of the factors have the superscript t. Next, we show that if $Z_C^0(Pa_G(C)) = 1$ then $Z_C^{t+1}(Pa_G(C)) = 1$. This implies that Equations 7 and 11 are equivalent because $Z_C^0(Pa_G(C)) = 1$ by line 3 and, thus, our assumption of Equation 11 is innocuous. To see it, let $U \in Cc(G)$, $Q \in Cs(G_U)$ and $W = C \setminus Q$. Then

$$p^{t+1}(Q|Pa_G(U)) = \sum_w p^{t+1}(w|Pa_G(U))$$

$$= \sum_w \psi_U^{t+1}(Q, Pa_G(Q)) \frac{1}{Z_U^{t+1}(Pa_G(U))} \prod_{K \in Cs(G_U) \setminus Q} \psi_U^t(k, Pa_G(K))$$

$$= \sum_w \psi_U^t(Q, Pa_G(Q)) \frac{p_e(Q|Pa_G(Q))}{p^t(Q|Pa_G(Q))} \frac{1}{Z_U^{t+1}(Pa_G(U))} \prod_{K \in Cs(G_U) \setminus Q} \psi_U^t(k, Pa_G(K))$$

$$= \frac{p_e(Q|Pa_G(Q))}{p^t(Q|Pa_G(Q))} \frac{Z_U^t(Pa_G(U))}{Z_U^{t+1}(Pa_G(U))} \sum_w p^t(w|Pa_G(U)) = \frac{p_e(Q|Pa_G(Q))}{p^t(Q|Pa_G(Q))} \frac{Z_U^t(Pa_G(U))}{Z_U^{t+1}(Pa_G(U))} p^t(Q|Pa_G(U))$$

$$= p_e(Q|Pa_G(Q)) \frac{Z_U^t(Pa_G(U))}{Z_U^{t+1}(Pa_G(U))}$$

since $p^t(Q|Pa_G(Q)) = p(Q|Pa_G(U))$ by C1* and decomposition. Summing both sides over q implies that the IPFP preserves the normalization constant across iterations.

3. Discussion

Given a probability distribution p that is Markovian wrt an AMP CG G, we have described in Equations 6-8 necessary and sufficient conditions for p to factorize wrt G. This note extends the original manuscript, where p is shown to factorize as

$$p(V) = \prod_{C \in Cc(G)} \prod_{K \in Cs(G_C)} \psi_C(K, Pa_G(K)).$$

To see that the condition above is necessary but not sufficient, consider the AMP CGs $A \rightarrow B - C$ and $A \rightarrow B - C \leftarrow A$, and note that both imply the same factorization, namely $p(A, B, C) = \psi_A(A)\psi_{BC}(A, B, C)$. So, if p encodes no independence then it factorizes according to both CGs although it is Markovian wrt only the second of them. In any case, the factorization above is enough to perform efficiently inference and parameter estimation, as shown in the original manuscript.

Unfortunately, finding the maximum likelihood estimates of the factors in the new factorization is difficult and, thus, we have decided to enforce only Equations 6 and 7 in the estimation process so that it can be performed efficiently via the IPFP. The so fitted factorization is enough to perform inference efficiently following the same procedure as in the original manuscript.

Our work is related to that by Drton (2008), where the author proposes necessary and sufficient conditions for p to factorize wrt G when G is a MVR CG. His factorization resembles ours in that it includes constraints similar to those in Equation 7, which make maximum likelihood estimation hard. To overcome this problem, the author develops a so called iterative conditional fitting procedure (ICFP) that, at each iteration, solves a convex optimization problem under the mentioned constraints. We plan to study whether it is possible to adapt the ICFP to our problem, given the similarity between the constraints in both factorizations. Drton (2008) also makes the interesting observation that the runtime of the ICFP can be shortened by replacing G with a Markov equivalent CG with smaller connectivity components. It would be interesting to see whether this also applies to our IPFP. A result that would be helpful in that investigation is that by Sonntag and Peña (2015, Theorem 4), which shows how to obtain a Markov equivalent CG with the fewest undirected edges.

Two other works that are related to ours are those by Abbeel et al. (2006) and Roy et al. (2009). Unlike our work, these works do not characterize when the Markovian and factorization properties are equivalent. Instead, they develop closed form expressions for estimating the factors in the factorization of p wrt G when G is a factor graph. Since factor graphs subsume AMP CGs, we can adapt their closed form estimates to our problem. Specifically, let $C \in Cc(G)$ and $K \in Cs(G_C)$. Also, let $Mb_G(K)$ denote the minimal subset of $CPa_G(C)$ st $K \perp_G CPa_G(C) \setminus KMb_G(K) | Mb_G(K)$. It is easy to see that $Mb_G(K) = Ne_G(K)Pa_G(K)Pa_G(Ne_G(K))$. Now, choose an arbitrary but fixed

state v^* of V. Then, we can use Proposition 4 by Abbeel et al. (2006) to rewrite the factorization in Corollary 1 as follows (details omitted):

$$p(v) = p(v^*) \prod_{C \in Cc(G)} \prod_{K \in Cs(G_C): K \neq \emptyset} \psi(k) \tag{12}$$

where k is consistent with v, and

$$\psi(k) = \exp\left(\prod_{B \subseteq K} (-1)^{|K \setminus B|} \log p(b, \overline{b}^* | mb_G(K)^*) \right) \tag{13}$$

where $\overline{b}^*$ and $mb_G(K)^*$ denote the values of $K \setminus B$ and $Mb(K)$ consistent with v^*. In order to estimate the factors above, the authors propose replacing p with the empirical probability distribution p_e. Unfortunately, this may produce an unreliable estimate for $p(v^*)$ unless the data available is abundant. Similarly for $p(b, \overline{b}^* | mb_G(K)^*)$ because K and $Mb_G(K)$ may be large. Note also that the estimate of $p(b, \overline{b}^* | mb_G(K)^*)$ is based only on the instances of the data that are consistent with $\overline{b}^*$ and $mb_G(K)^*$ simultaneously. This means that the data available is not used efficiently. All this leads the authors to acknowledge that their closed form estimates, as described, are probably impractical (Abbeel et al., 2006, p. 1764). Roy et al. (2009) improve the method above by simplifying Equation 13. Although the improvement alleviates the drawbacks mentioned, it does not eliminate them completely (e.g. Equation 12 stays the same and, thus, the problem of estimating $p(v^*)$ remains). Unfortunately, no experimental results are reported in either of the works cited. It would be interesting to compare them with our IPFP.

References

Abbeel, P., Koller, D. and Ng, A. Y. Learning Factor Graphs in Polynomial Time and Sample Complexity. *Journal of Machine Learning Research*, 7:1743-1788, 2006.

Andersson, S. A., Madigan, D. and Perlman, M. D. Alternative Markov Properties for Chain Graphs. *Scandinavian Journal of Statistics*, 28:33-85, 2001.

Drton, M. Iterative Conditional Fitting for Discrete Chain Graph Models. In *Proceedings in Computational Statistics*, 93-104, 2008.

Lauritzen, S. L. *Graphical Models*. Oxford University Press, 1996.

Murphy, K. P. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

Roy, S., Lane, T. and Werner-Washburne, M. Learning Structurally Consistent Undirected Probabilistic Graphical Models. In *Proceedings of the 26th International Conference on Machine Learning*, 905-912, 2009.

Sonntag, D. and Peña, J. M. Chain Graph Interpretations and their Relations Revisited. *International Journal of Approximate Reasoning*, 58:3956, 2015.