Look Ma, No Latent Variables: Accurate Cutset Networks via Compilation Supplementary Material

Tahrima Rahman [*1] **Shasha Jin** [*1] **Vibhav Gogate** [1]

Proof of Theorem 1

Proof. We prove this theorem by induction. Base case is trivially true.

In the inductive step, we assume that the theorem is true for depth d and prove that it also holds for depth $d+1$.

Let $\mathcal{C}$ be a cutset network having depth d and representing a probability distribution R_d. Without loss of generality, we assume that the OR tree in $\mathcal{C}$ is balanced and complete. Let $\mathcal{C}'$ be a cutset network representing a distribution R_{d+1} having depth $d+1$ constructed from $\mathcal{C}$ as follows. Pick an arbitrary leaf node l of $\mathcal{C}$. Let T_l be the tree Bayesian network at l representing the distribution P_l. Replace T_l by the following cutset network stump (we call a cutset network having just one OR node and two tree Bayesian networks as leaf nodes of the OR node as a cutset network stump). Pick an arbitrary variable, say X in T_l as the root node of the cutset network stump. We consider two variations of the cutset network stump.

- **Variation 1**: Tree Bayesian networks attached to each of the two leaf nodes of the cutset network stump are learned using the Chow-Liu algorithm with the pairwise mutual information scores over all pairs of variables in $V(T_l) \setminus \{X\}$ computed from the latent tractable model Q. Labels on the edges in the OR tree, namely $P(X=0)$ and $P(X=1)$ are computed by performing inference over Q. Let $P^{(1)}$ be the distribution represented by the resulting cutset network stump.
- **Variation 2**: Tree Bayesian networks attached to each of the two leaf nodes of the cutset network stump are constructed using the following method. Tree Bayesian network at the left leaf and right leaf is constructed by setting $X=0$ and $X=1$ as evidence respectively in T_l and normalizing. Labels on the edges in the OR tree, namely $P(X=0)$ and $P(X=1)$ are computed by performing inference over T_l. Let $P^{(2)}$ be the distribution represented by the resulting cutset network stump.

By construction, Variation 2 yields a cutset network stump having the same distribution as T_l, namely $P^{(2)} = P_l$. Moreover, since Variation 1 uses the Chow-Liu algorithm and the latter is an optimal algorithm, each tree Bayesian network constructed using Variation 1 is superior (or the same) as the one using Variation 2 in terms of KL divergence computed with respect to $Q(\mathbf{x}_{V(T_l)}|\mathbf{x}_{path(l)})$. In other words, the KL divergence between $Q(\mathbf{x}_{V(T_l)}|\mathbf{x}_{path(l)})$ and $P^{(1)}$ is smaller than or equal to the KL divergence between $Q(\mathbf{x}_{V(T_l)}|\mathbf{x}_{path(l)})$ and $P^{(2)} = P_l$. Therefore, it follows that the KL divergence between Q and R_{d+1} is smaller than or equal to the one between Q and R_d. Note that since X was chosen arbitrarily, it applies to any variable chosen (whether chosen heuristically or not) by Algorithm 1. This proves the inductive step and completes the proof. $\square$

Additional Experimental Results

Density Estimation

In addition to algorithms CNxD and CN described in the main text of the paper, in the supplement we consider an additional variation CNx obtained via setting α to 1 in Eq. (4). We call it **CNx** which stands for <u>c</u>utset <u>n</u>etworks learned using e<u>x</u>act inference performed on the latent tractable model Q (and data is not used to compute the sufficient statistics). Figures 1 and 2 show the test set log-likelihood scores of CNxDs, CNxs and CNRs as a function of time for all the 20 datasets. We see that CNRs quickly reach a reasonable solution but are unable to improve their solution at the same rate as CNxDs and CNxs. In general, CNxDs are almost always better than CNxs, and when the time bound is sufficiently large, CNxDs outperform CNxs, CNRs and CNs and their performance approaches that of latent tractable models (BCNs and MTs).

[*]Equal contribution [1]Department of Computer Science, The University of Texas at Dallas, United States. Correspondence to: Tahrima Rahman <tahrima.rahman@utdallas.edu>.

Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

Table 1. Average test set log-likelihood scores of CNxDs, CNxs, CNs and CNRs.

Dataset	Average test set log-likelihood			
	CNxD	CNx	CN	CNR
	$\alpha \in [0, 1]$	$\alpha = 1$	$\alpha = 0$	$\alpha = 1$
nltcs	-6.01	-6.02	-6.05	**-5.97**
msnbc	-6.07	-6.08	-6.05	**-6.03**
kdd	**-2.15**	**-2.15**	-2.19	-2.16
plants	**-12.73**	-12.96	-13.25	-15.00
audio	**-40.69**	-40.82	-41.97	-41.97
jester	**-53.67**	-53.83	-55.26	-54.66
netflix	**-57.48**	-57.63	-58.72	-59.15
accidents	**-30.12**	-30.38	-30.66	-38.54
retail	**-10.84**	-10.88	-10.98	-11.27
pumsb*	**-23.57**	-23.78	-24.28	-36.16
dna	-87.98	-88.59	**-87.50**	-96.63
kosarek	**-10.74**	-10.77	-11.07	-11.97
msweb	**-9.76**	-9.82	-10.12	-11.12
book	**-35.31**	-35.81	-37.51	-37.22
movie	**-54.61**	-55.42	-57.71	-65.95
webkb	**-155.77**	-157.98	-161.58	-172.13
reuters	**-85.89**	-86.55	-87.64	-101.16
20newsg	**-155.66**	-156.71	-161.68	-164.34
bbc	**-253.50**	-260.65	-260.55	-271.98
ad	**-15.40**	-16.36	-16.14	-52.74
Average	**-55.40**	-56.16	-57.05	-62.81

Prediction Accuracy: MAP Inference

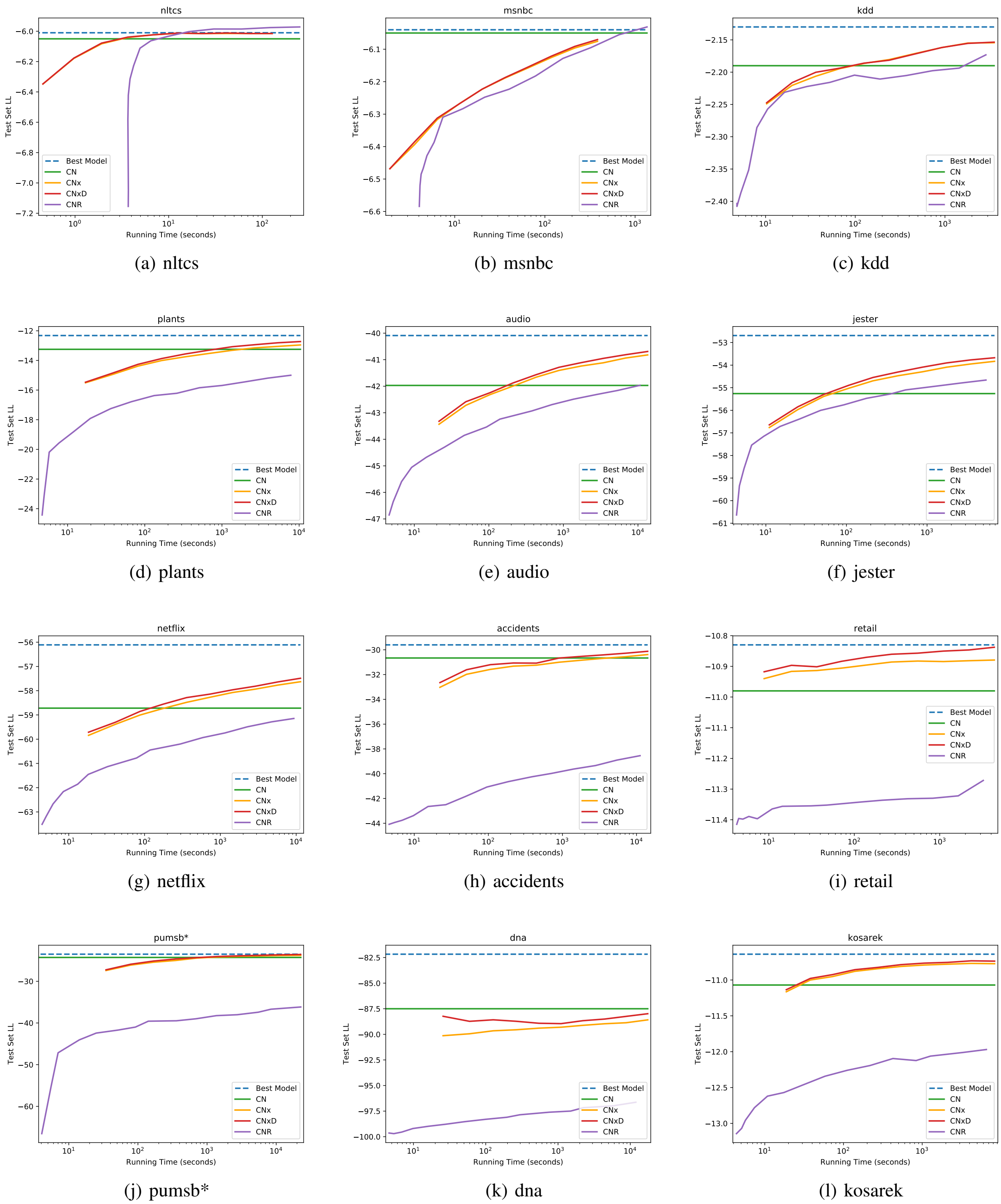
Tables 2 and 3 as well as Figures 3–7 compare the quality of MAP estimates output by various algorithms using the Hamming Loss and F1 score criteria respectively. We observe a similar trend to the log-likelihood criteria reported in the main paper: CNxDs dominate latent models (BCNs and MTs) as well as CNs (which are learned from data alone). CNRs is the worst performing method which shows the utility of structure learning.

Table 2. Hamming loss Comparison (Quality of MAP inference). Bold values indicate best scores obtained by CN, CNxD, CNR, MT or BCN.

Datasets	20% Evidence					50% Evidence					80% Evidence				
	MAP Tractable			MAP Intractable		MAP Tractable			MAP Intractable		MAP Tractable			MAP Intractable	
	CNxD	CN	CNR	MT	BCN	CNxD	CN	CNR	MT	BCN	CNxD	CN	CNR	MT	BCN
nltcs	**0.1998**	0.2116	0.2032	0.2173	0.2220	**0.1417**	0.1615	0.1707	0.1647	0.1644	**0.1270**	0.1321	0.0957	0.1364	0.1300
msnbc	**0.1636**	0.1849	0.1817	0.1979	0.1850	0.1636	0.1636	**0.1349**	0.1638	0.1638	0.1443	0.1448	**0.1399**	0.1456	0.1528
kdd	0.0064	0.0064	**0.0057**	0.0064	0.0065	**0.0055**	0.0065	0.0069	0.0064	0.0065	**0.0063**	0.0069	0.0065	0.0065	0.0067
plants	**0.0903**	0.1018	0.1007	0.0990	0.1022	**0.0663**	0.0710	0.0714	0.0713	0.0720	**0.0495**	0.0629	0.0693	0.0627	0.0630
audio	0.1916	0.1921	0.1922	**0.1894**	0.1899	0.1766	0.1796	**0.1762**	0.1783	0.1791	**0.1682**	0.1733	0.1686	0.1722	0.1724
jester	0.3149	0.3160	0.3162	0.3160	**0.3135**	**0.2664**	0.2730	0.2710	0.2705	0.2724	**0.2602**	0.2724	0.2747	0.2639	0.2678
netflix	**0.3478**	0.3526	0.3888	0.3849	0.3704	**0.3028**	0.3093	0.3033	0.3068	0.3065	**0.2776**	0.2873	0.2884	0.2820	0.2813
accidents	**0.1501**	0.1622	0.1811	0.1664	0.1663	**0.0939**	0.1071	0.0999	0.1043	0.1059	**0.0795**	0.0945	0.0842	0.0902	0.0916
retail	**0.0210**	0.0231	0.0250	0.0232	0.0232	**0.0162**	0.0217	0.0320	0.0219	0.0221	**0.0115**	0.0136	0.0116	0.0137	0.0141
pumsb*	**0.0804**	0.0844	0.0832	0.0849	0.0841	**0.0459**	0.0499	0.0497	0.0489	0.0497	**0.0414**	0.0429	0.0477	0.0416	0.0427
dna	**0.3056**	0.3239	0.3300	0.3421	0.3299	**0.2466**	0.2702	0.2675	0.2818	0.2813	**0.2358**	0.2471	0.2391	0.2379	0.2439
kosarek	**0.0160**	0.0185	0.0183	0.0182	0.0183	**0.0112**	0.0170	0.0181	0.0168	0.0170	0.0199	0.0202	**0.0149**	0.0197	0.0202
msweb	0.0114	**0.0103**	0.0115	0.0115	0.0120	**0.0080**	0.0106	0.0143	0.0114	0.0114	**0.0102**	0.0106	0.0184	0.0108	0.0109
book	**0.0163**	0.0166	0.0166	0.0166	0.0166	**0.0166**	0.0170	0.0168	0.0168	0.0170	**0.0152**	0.0174	0.0155	0.0170	0.0173
movie	**0.0453**	0.0478	0.0498	0.0487	0.0484	**0.0388**	0.0438	0.0434	0.0435	0.0434	**0.0421**	0.0437	0.0438	0.0423	0.0426
webkb	**0.0636**	0.0643	0.0649	0.0645	0.0644	0.0609	0.0614	**0.0584**	0.0610	0.0608	**0.0570**	0.0577	0.0645	0.0572	0.0574
reuters	**0.0314**	0.0318	0.0319	0.0319	0.0323	**0.0284**	0.0291	0.0291	0.0290	0.0289	**0.0269**	0.0287	0.0259	0.0282	0.0280
20newsg	**0.0506**	0.0528	0.0528	0.0522	0.0525	**0.0493**	0.0514	0.0504	0.0509	0.0513	**0.0504**	0.0513	0.0515	0.0502	0.0510
bbc	**0.0791**	0.0798	0.0799	0.0801	0.0804	0.0778	0.0787	**0.0761**	0.0785	0.0769	0.0755	0.0774	0.0796	0.0769	**0.0742**
ad	**0.0027**	0.0029	0.0029	0.0028	0.0030	**0.0013**	0.0015	0.0014	0.0015	0.0015	**0.0009**	0.0012	0.0011	0.0012	0.0011
Average Loss	**0.1094**	0.1142	0.1168	0.1177	0.1160	**0.0909**	0.0962	0.0946	0.0964	0.0966	**0.0850**	0.0893	0.0870	0.0878	0.0885
Wins/Total	**16/20**	1/20	1/20	1/20	1/20	**16/20**	0/20	4/20	0/20	0/20	**17/20**	0/20	2/20	0/20	1/20

Table 3. F1 Score Comparison (Quality of MAP inference). Bold values indicate highest score obtained by a model.

Datasets	20% Evidence					50% Evidence					80% Evidence				
	MAP Tractable			MAP Intractable		MAP Tractable			MAP Intractable		MAP Tractable			MAP Intractable	
	CNxD	CN	CNR	MT	BCN	CNxD	CN	CNR	MT	BCN	CNxD	CN	CNR	MT	BCN
nltcs	**0.6756**	0.2029	0.6686	0.2066	0.2128	**0.6712**	0.1435	0.6704	0.1439	0.1466	**0.7541**	0.1684	0.7508	0.1718	0.1646
msnbc	0.1442	0.1812	0.1084	**0.199**	0.1811	**0.2282**	0.1807	0.1832	0.1801	0.1824	**0.512**	0.2382	0.4242	0.2408	0.242
kdd	**0.1089**	0.0065	0.0842	0.0065	0.0066	**0.1483**	0.0057	0.1175	0.0056	0.0056	**0.0993**	0.007	0.0721	0.0065	0.0066
plants	**0.6768**	0.0966	0.6117	0.0945	0.0976	**0.7691**	0.0683	0.6811	0.0702	0.07	**0.7695**	0.0535	0.6707	0.0506	0.052
audio	0.1011	**0.1924**	0.0359	0.1883	0.1887	**0.3395**	0.1786	0.2425	0.1784	0.1794	**0.4015**	0.1793	0.2926	0.1776	0.1779
jester	**0.7606**	0.3267	0.7457	0.3214	0.3196	**0.7562**	0.2716	0.7561	0.2685	0.2701	**0.7923**	0.2548	0.7837	0.2452	0.2534
netflix	0.6538	0.3555	**0.6654**	0.4001	0.3753	**0.6814**	0.3231	0.6634	0.3197	0.3194	**0.7268**	0.2862	0.6931	0.2814	0.281
accidents	**0.3927**	0.1597	0.3166	0.1586	0.1612	**0.4761**	0.0981	0.3281	0.0927	0.0939	**0.6148**	0.0845	0.3568	0.0765	0.0815
retail	0.0066	0.0211	0.0004	0.0212	0.0211	**0.0109**	0.0163	0.0000	0.0164	0.0171	**0.0169**	0.0116	0.0007	0.0117	0.0118
pumsb*	**0.6252**	0.0842	0.4941	0.0836	0.0827	**0.6859**	0.0475	0.5041	0.047	0.0479	**0.7979**	0.0572	0.5231	0.055	0.0564
dna	0.2402	0.3261	0.1413	**0.344**	0.3314	0.219	0.2659	0.0643	0.2779	**0.278**	0.195	**0.2413**	0.0522	0.2399	0.241
kosarek	**0.0372**	0.0204	0.012	0.0201	0.02	**0.0811**	0.0113	0.0245	0.0112	0.0113	**0.1338**	0.0242	0.0674	0.024	0.0244
msweb	0.0094	0.0115	0.0009	**0.013**	0.0136	**0.0402**	0.0081	0.0023	0.0084	0.0085	**0.0636**	0.0102	0.0001	0.0103	0.0104
book	**0.0235**	0.0164	0.0051	0.0164	0.0164	**0.0587**	0.0179	0.0082	0.0176	0.0179	**0.0663**	0.016	0.0209	0.0157	0.016
movie	**0.0938**	0.0446	0.0284	0.0468	0.0462	**0.1418**	0.0395	0.072	0.0392	0.0395	**0.2699**	0.0521	0.1346	0.0513	0.051
webkb	0.0335	0.0645	0.0073	**0.0647**	0.0646	**0.0865**	0.0633	0.0158	0.0631	0.0629	**0.1333**	0.0602	0.0233	0.0595	0.0596
reuters	0.0293	**0.0315**	0.0028	0.0312	0.0314	**0.0749**	0.0305	0.0088	0.0307	0.0303	**0.1046**	0.0284	0.0138	0.0273	0.0274
20newsg	0.0136	**0.0528**	0.0057	0.0523	0.0527	0.0333	**0.0497**	0.0069	0.0496	0.0496	**0.075**	0.0543	0.0253	0.0533	0.0543
bbc	0.0145	0.0793	0.0021	**0.0794**	0.08	0.067	**0.0793**	0.005	0.0792	0.0778	**0.1035**	0.0783	0.0061	0.0773	0.075
ad	**0.5774**	0.0029	0.035	0.0029	0.003	**0.7548**	0.0018	0.1384	0.0018	0.0018	**0.7843**	0.001	0.0931	0.0011	0.001
Average F1 Score	**0.2609**	0.1138	0.1986	0.1175	0.1153	**0.3162**	0.0950	0.2246	0.0951	0.0955	**0.3707**	0.0953	0.2502	0.0938	0.0944
Wins/Total	**11/20**	3/20	1/20	5/20	0/20	**17/20**	2/20	0/20	0/20	1/20	**19/20**	1/20	0/20	0/20	0/20



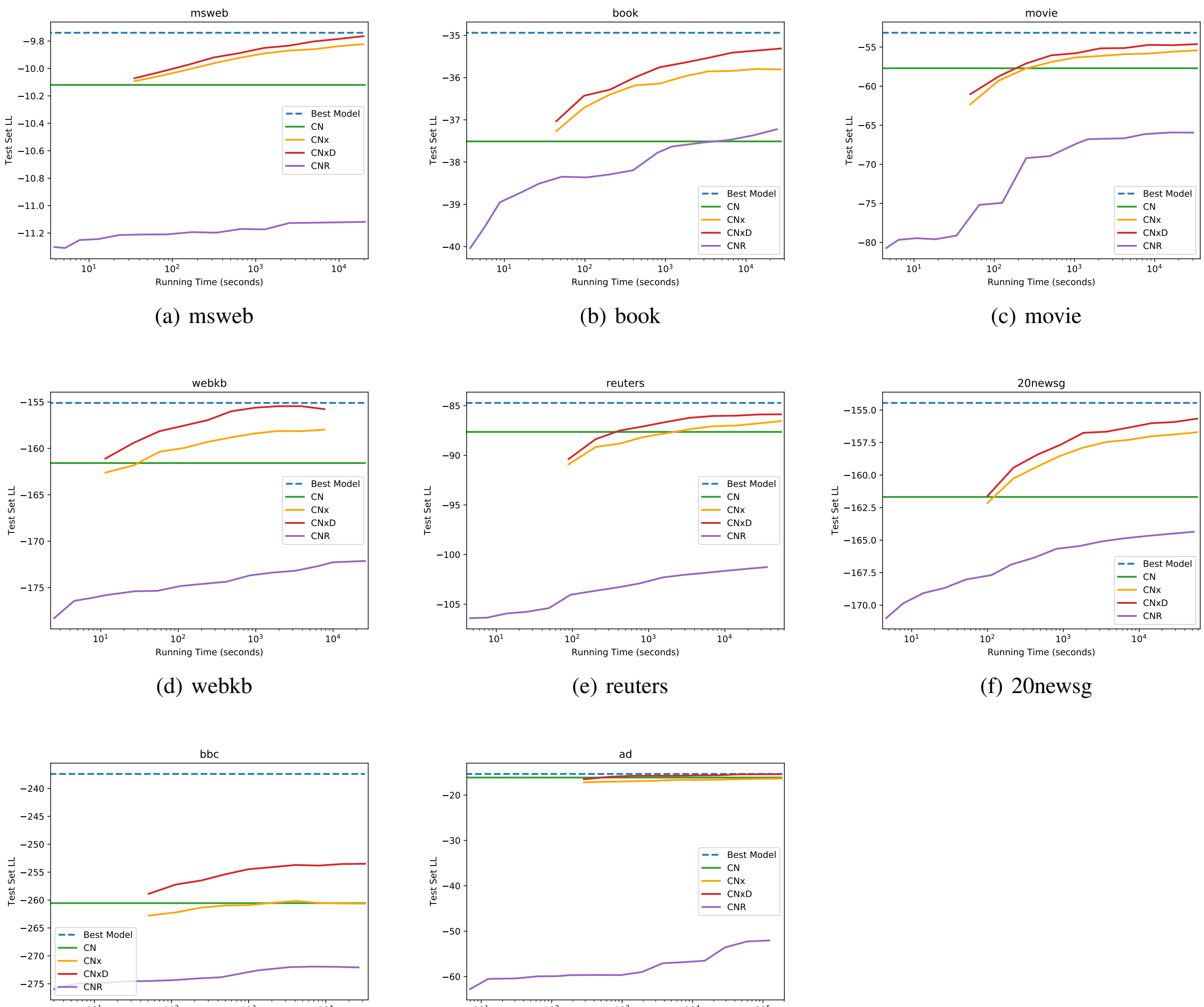
(a) nltcs (b) msnbc (c) kdd

(d) plants (e) audio (f) jester

(g) netflix (h) accidents (i) retail

(j) pumsb* (k) dna (l) kosarek

Figure 1. Average test set log-likelihood as a function of the running time on the first 12 datasets.

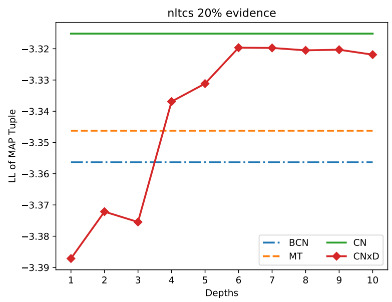


(a) msweb (b) book (c) movie

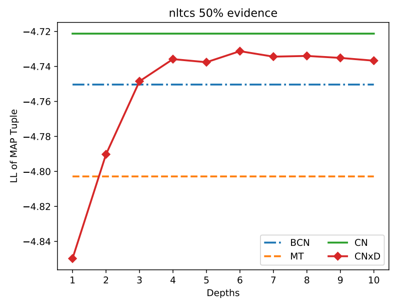
(d) webkb (e) reuters (f) 20newsg

(g) bbc (h) ad

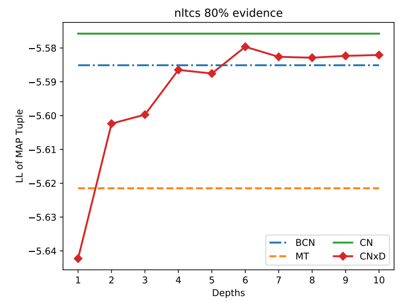
Figure 2. Average test set log-likelihood as a function of the running time on the last 8 datasets.



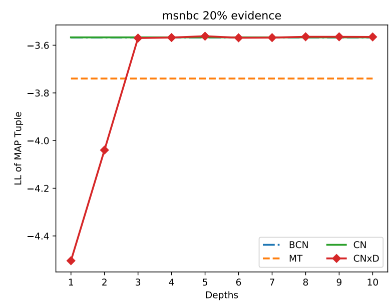
(a) 20% Evidence; nltcs



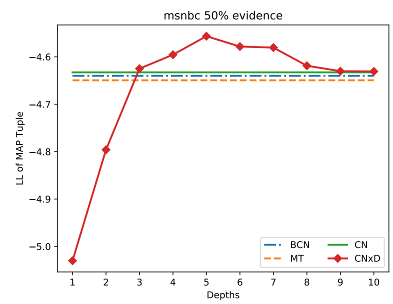
(b) 50% Evidence; nltcs



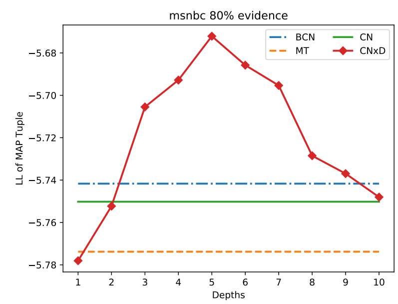
(c) 80% Evidence; nltcs



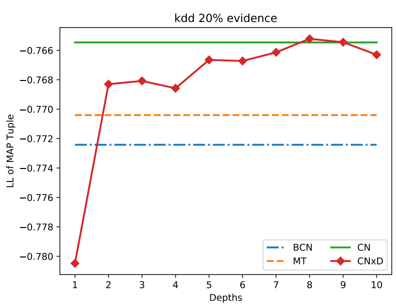
(d) 20% Evidence; msnbc



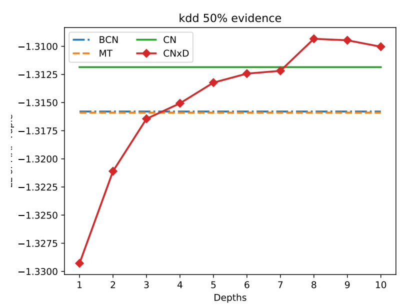
(e) 50% Evidence; msnbc



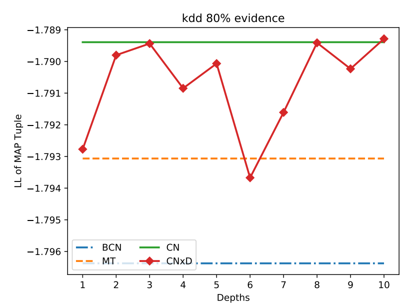
(f) 80% Evidence; msnbc



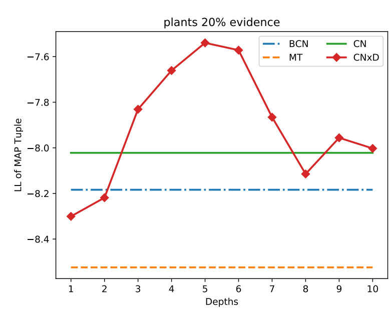
(g) 20% Evidence; kdd



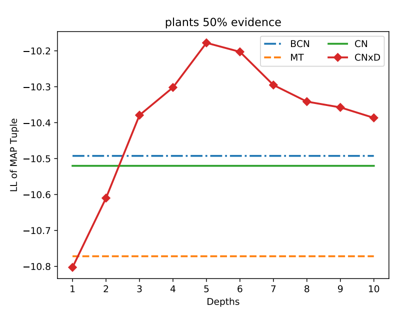
(h) 50% Evidence; kdd



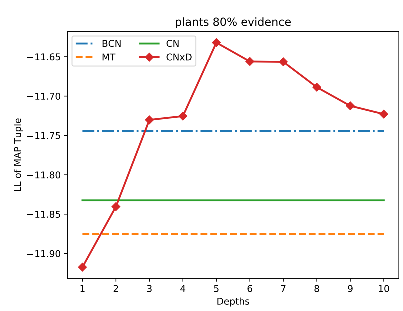
(i) 80% Evidence; kdd



(j) 20% Evidence; plants

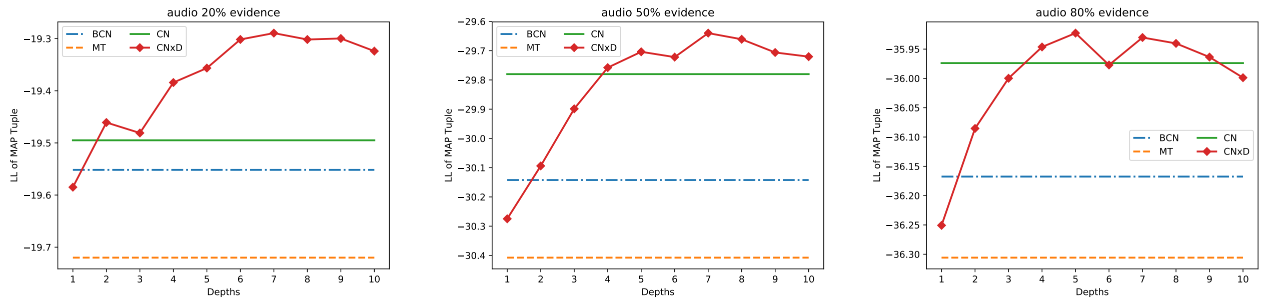


(k) 50% Evidence; plants

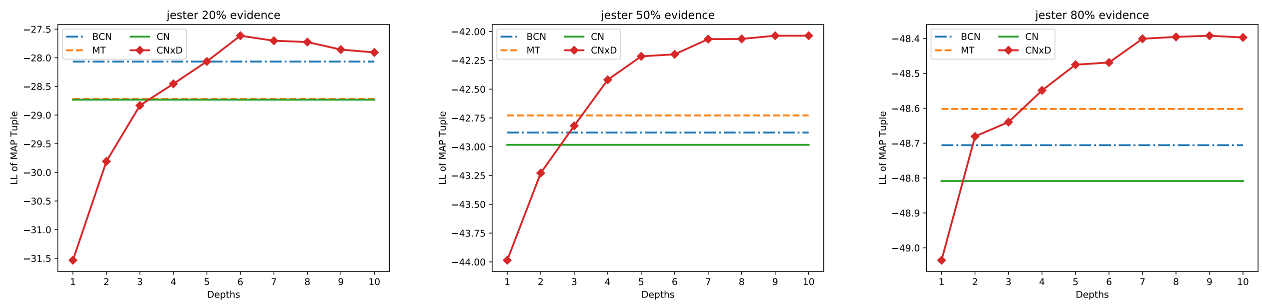


(l) 80% Evidence; plants

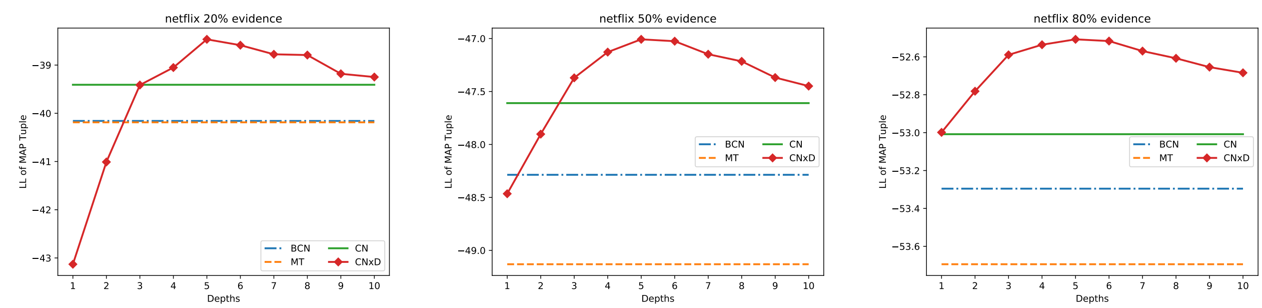
Figure 3. Average test set log-likelihood score of the MAP completion of evidence output by various algorithms as a function of the depth of the model on datasets nltcs, msnbc, kdd, plants.



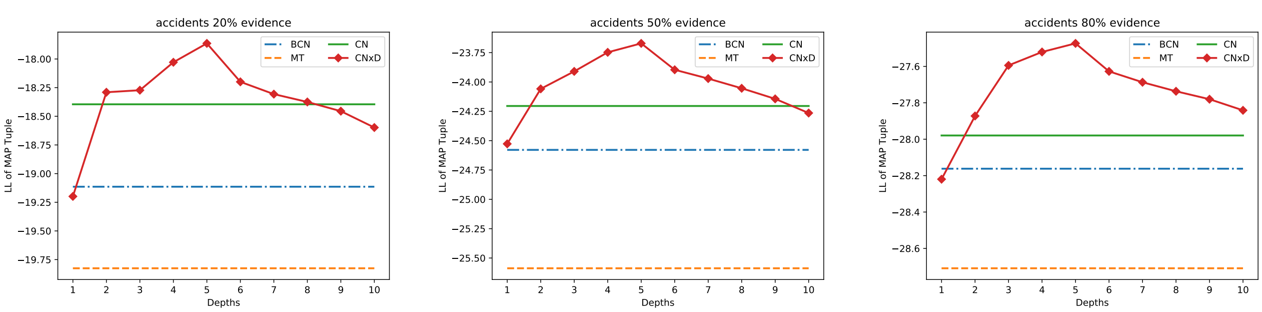
(a) 20% Evidence; audio (b) 50% Evidence; audio (c) 80% Evidence; audio



(d) 20% Evidence; jester (e) 50% Evidence; jester (f) 80% Evidence; jester

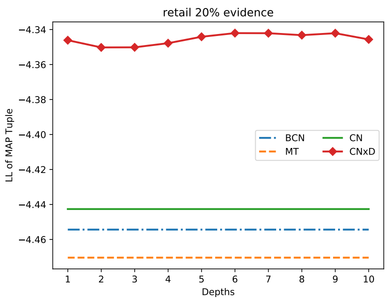


(g) 20% Evidence; netflix (h) 50% Evidence; netflis (i) 80% Evidence; netflix

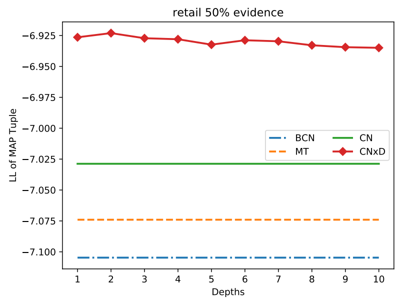


(j) 20% Evidence; accidents (k) 50% Evidence; accidents (l) 80% Evidence; accidents

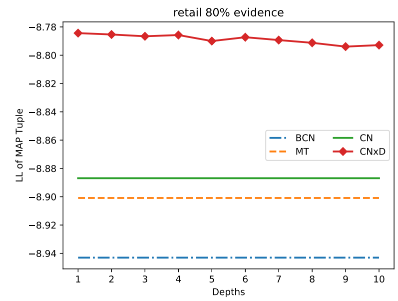
Figure 4. Average test set log-likelihood score of the MAP completion of evidence output by various algorithms as a function of the depth of the model on datasets audio, jester, netflix, accidents.



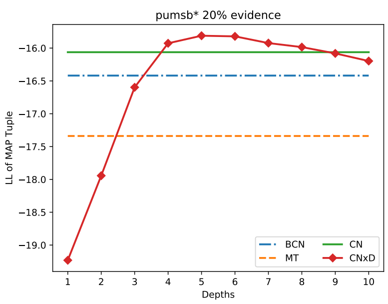
(a) 20% Evidence; retail



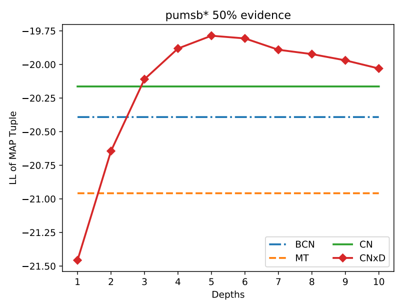
(b) 50% Evidence; retail



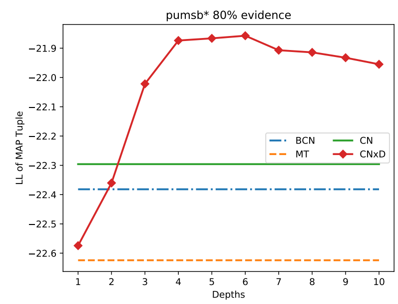
(c) 80% Evidence; retail



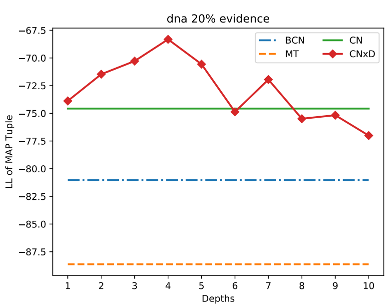
(d) 20% Evidence; pumsb*



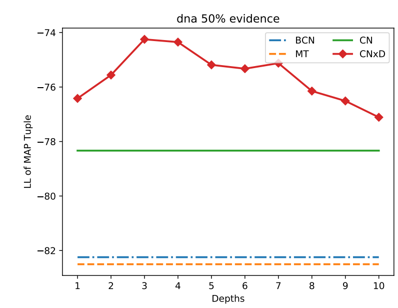
(e) 50% Evidence; pumsb*



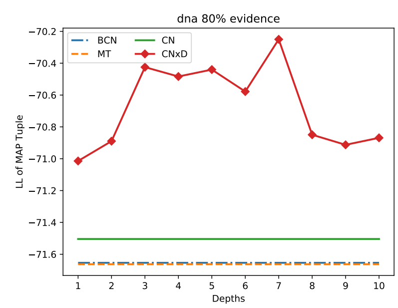
(f) 80% Evidence; pumsb*



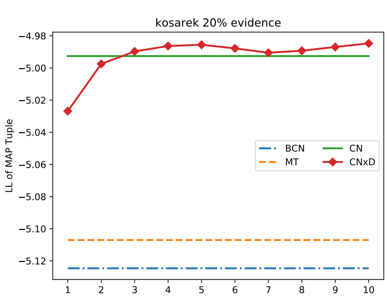
(g) 20% Evidence; dna



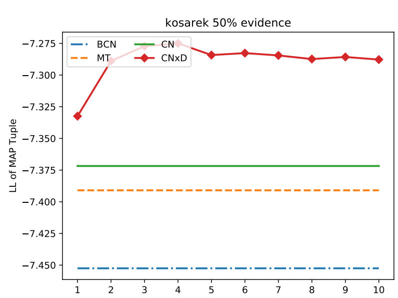
(h) 50% Evidence; dna



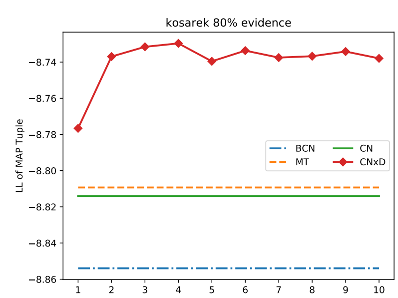
(i) 80% Evidence; dna



(j) 20% Evidence; kosarek

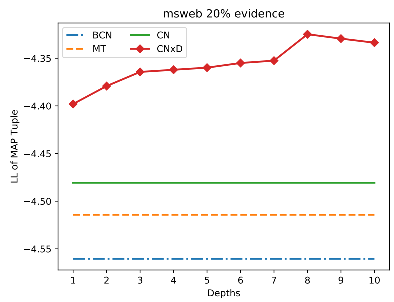


(k) 50% Evidence; kosarek

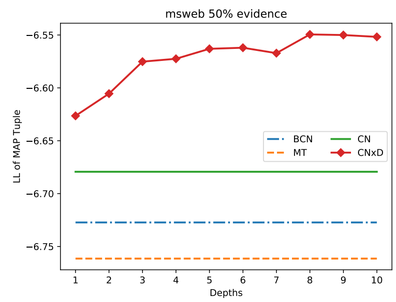


(l) 80% Evidence; kosarek

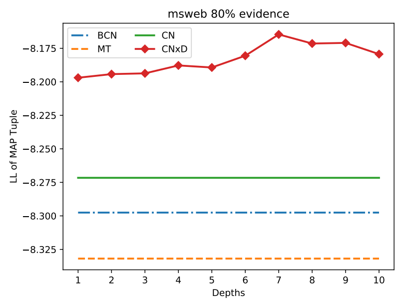
Figure 5. Average test set log-likelihood score of the MAP completion of evidence output by various algorithms as a function of the depth of the model on datasets retail, pumsb*, dna, kosarek.



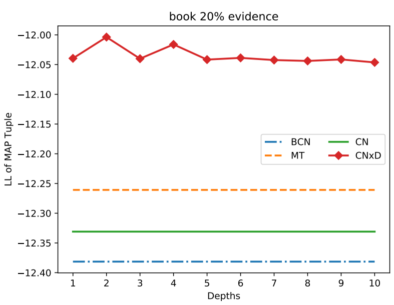
(a) 20% Evidence; msweb



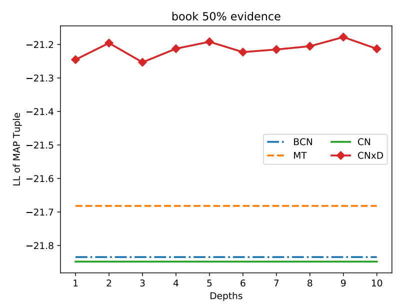
(b) 50% Evidence; msweb



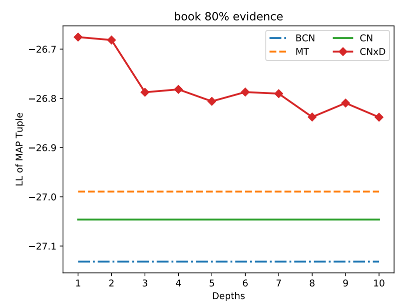
(c) 80% Evidence; msweb



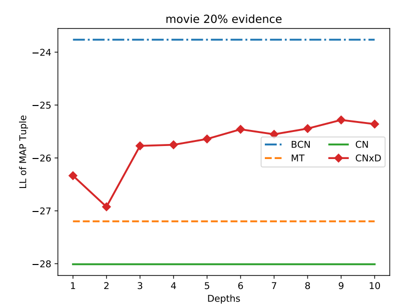
(d) 20% Evidence; book



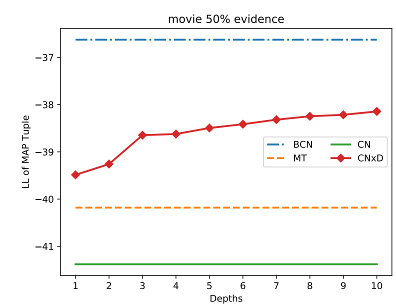
(e) 50% Evidence; book



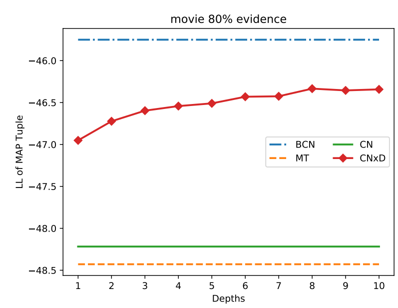
(f) 80% Evidence; book



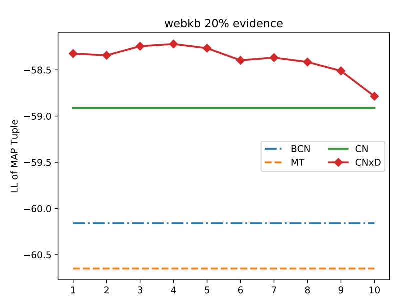
(g) 20% Evidence; movie



(h) 50% Evidence; movie

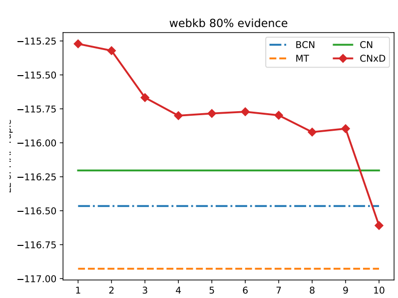


(i) 80% Evidence; movie



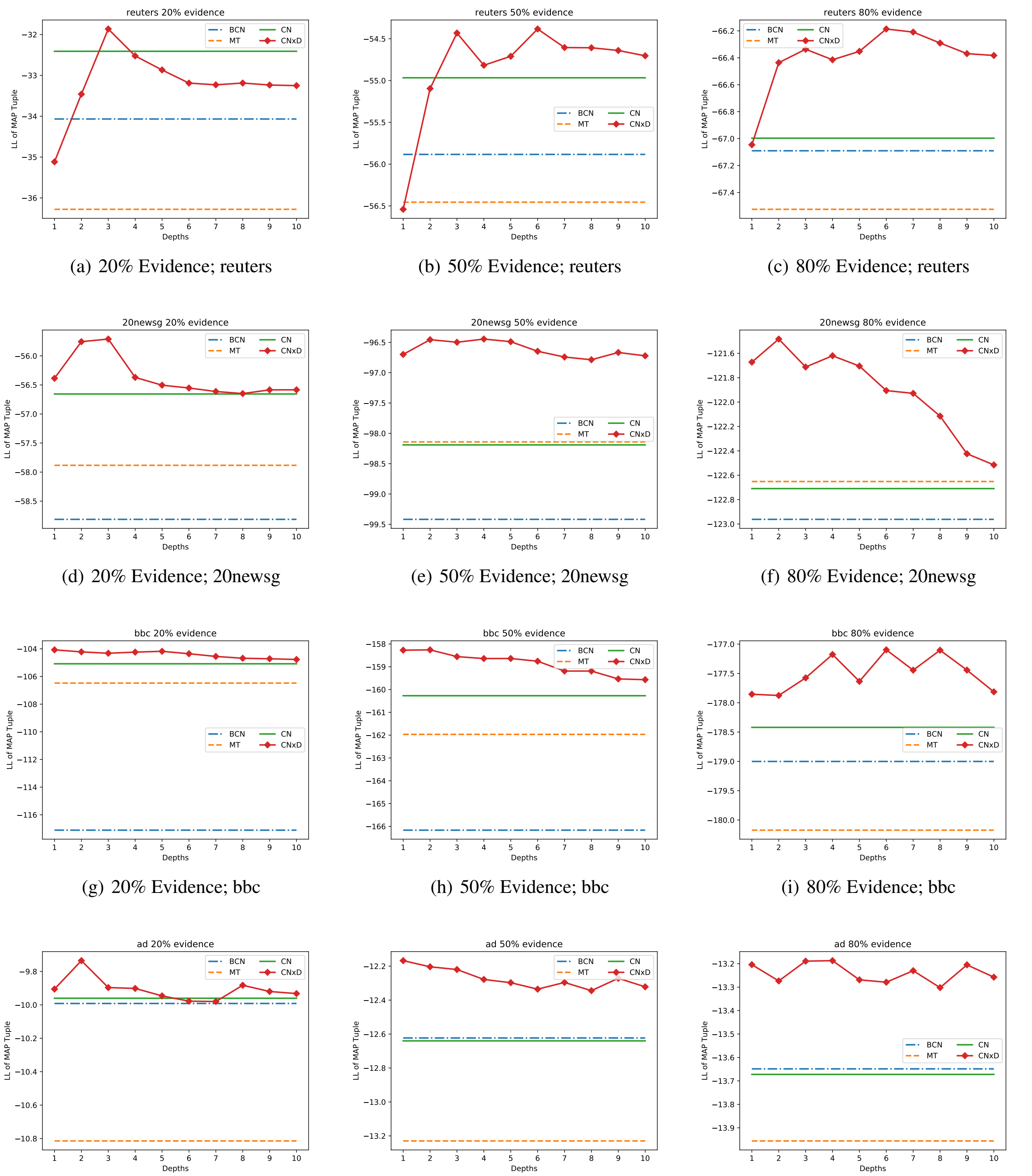
(j) 20% Evidence; webkb

(k) 50% Evidence; webkb



(l) 80% Evidence; webkb

Figure 6. Average test set log-likelihood score of the MAP completion of evidence output by various algorithms as a function of the depth of the model on on datasets msweb, book, movie, webkb.



(a) 20% Evidence; reuters (b) 50% Evidence; reuters (c) 80% Evidence; reuters

(d) 20% Evidence; 20newsg (e) 50% Evidence; 20newsg (f) 80% Evidence; 20newsg

(g) 20% Evidence; bbc (h) 50% Evidence; bbc (i) 80% Evidence; bbc

(j) 20% Evidence; ad (k) 50% Evidence; ad (l) 80% Evidence; ad

Figure 7. Average test set log-likelihood score of the MAP completion of evidence output by various algorithms as a function of the depth of the model on datasets reuters, 20newsg, bbc, ad.