NeuroBE: Escalating Neural Network Approximations of Bucket Elimination (Supplementary material)

Sakshi Agarwal[1] **Kalev Kask**[1] **Alex Ihler**[1] **Rina Dechter**[1]

[1]University of California Irvine

0.1 ESTIMATING THE PSEUDO-DIMENSION OF A NN:

In our work, we use NN architectures with ReLU activation functions. To construct a NN with L layers and a variable h, #hidden-units per layer to model a specific local bucket message μ^*, we pick the rule $h = b * w$ where w is the bucket's width and b is a constant. By doing this, the #parameters in the NN is :

$$|\theta| = (L-1) * b^2 * w^2 + b * w^2 + (L+1) * b * w + 1 \quad (1)$$

We make use of the lower bound of pseudo-dimension for NNs with ReLU activation functions from the work in Bartlett et al. [2019] to get:

$$\rho = |\theta| * Llog(|\theta|/L) \quad (2)$$

By substituting Eq. 1 in Eq. 2 and ignoring all linear terms in w we get that ρ can be dominated by:

$$\rightarrow \rho \propto (L * b * w)^2 log[(b * w)]$$

0.2 ESTIMATING ERROR IN PARTITION FUNCTION:

Let B_c be a bucket in a bucket chain along an ordering d; let B_c contain the original functions as ϕ_c and μ_{c+1} as the message passed to it from the previous bucket; let λ_c be the (global) exact message generated in B_c, μ_c^* be the local exact message in B_c and $\mu_c = APP(\mu_c^*)$ its approximation (e.g., by a trained neural network). Let $E_c = \log \mu_c^* - \log \mu_c$ and $\epsilon_c = max_{B_c}|E_c|$. Then,

$$\log \lambda_c - \log \mu_c \leq \sum_{k=2}^{n-c} \epsilon_{c+k}$$

In particular, since $\lambda_1 = Z$, the partition function and $\mu_1 = \hat{Z}$, the estimate to the partition function,

$$\log Z - \log \hat{Z} \leq \sum_{k=1}^{n-1} \epsilon_{1+k} \quad (3)$$

We will next derive the recursion, starting at the first processed bucket B_n and going down in order. Remember throughout that $\log \mu_{n-i}^* = \log \sum_{X_{n-i}} (e^{\log \phi_{n-i} + \log \mu_{n-i+1}})$

For B_n $\lambda_n = \mu_n^*$, therefore

$$\log \lambda_n - \log \mu_n = \log \mu_n^* - \log \mu_n = E_n$$

For B_{n-1}, by definition

$$\log \lambda_{n-1} - \log \mu_{n-1} = \log \sum_{X_{n-1}} e^{\log \phi_{n-1} + \log \lambda_n} - \log \mu_{n-1}$$

Substituting $\log \lambda_n$ from B_n

$$= \log \sum_{X_{n-1}} e^{[(\log \phi_{n-1} + \log \mu_n) + E_n]} - \log \mu_{n-1}$$

$$= \log[\sum_{X_{n-1}} e^{(\log \phi_{n-1} + \log \mu_n)} e^{E_n}] - \log \mu_{n-1}$$

If $max_{scope(\mu_n^*)}|E_n| = \epsilon_n$, then,

$$\leq \log[e^{\epsilon_n} \sum_{X_{n-1}} e^{(\log \phi_{n-1} + \log \mu_n)}] - \log \mu_{n-1}$$

$$\leq \epsilon_n + \log \sum_{X_{n-1}} e^{(\log \phi_{n-1} + \log \mu_n)} - \log \mu_{n-1}$$

Since $\log \sum_{X_{n-1}} e^{\log \phi_{n-1} + \log \mu_n} = \log \mu_{n-1}^*$ we get

$$\log \lambda_{n-1} - \log \mu_{n-1} \leq \epsilon_n + \log \mu_{n-1}^* - \log \mu_{n-1} \quad (4)$$

Accepted for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).

or equivalently,

$$\log \lambda_{n-1} - \log \mu_{n-1} \leq \epsilon_n + E_{n-1} \tag{5}$$

Moving to B_{n-2}, by definition:

$$\log \lambda_{n-2} - \log \mu_{n-2} = \log \sum_{X_{n-2}} e^{\log \phi_{n-2} + \log \lambda_{n-1} - \log \mu_{n-2}} \tag{6}$$

Substituting $\log \lambda_{n-1}$ from Eq. (4) we get

$$\log \lambda_{n-2} - \log \mu_{n-2}$$

$$\leq \log \sum_{X_{n-2}} e^{\log \phi_{n-2} + [\log \mu_{n-1} + \epsilon_n + E_{n-1}]} - \log \mu_{n-2}$$

$$\leq \log \sum_{X_{n-2}} e^{\log \phi_{n-2} + \mu_{n-1}} e^{\epsilon_n + E_{n-1}} - \log \mu_{n-2}$$

Taking $\max_{scope(\mu^*_{n-1})} E_{n-1} = \epsilon_{n-1}$,

$$\leq^{\epsilon_n + \epsilon_{n-1}} \sum_{X_{n-2}} e^{\log \phi_{n-2} + \mu_{n-1}} - \log \mu_{n-2}$$

$$\leq \epsilon_n + \epsilon_{n-1} + \log \sum_{X_{n-2}} e^{\log \phi_{n-2} + \mu_{n-1}} - \log \mu_{n-2}$$

$$\leq \epsilon_n + \epsilon_{n-1} + \log \mu^*_{n-2} - \log \mu_{n-2}$$

yielding,

$$\log \lambda_{n-2} - \log \mu_{n-2} \leq E_{n-2} + \epsilon_{n-1} + \epsilon_n \tag{7}$$

Moving to bucket B_{n-3}, by definition

$$\log \lambda_{n-3} - \log \mu_{n-3} = \log \sum_{X_{n-3}} e^{\log \phi_{n-3} + \log \lambda_{n-2} - \log \mu_{n-3}}$$

Substituting for λ_{n-2} from Eq. (7) we get with some algebra

$$\log \lambda_{n-3} - \log \mu_{n-3}$$

$$\leq \log \sum_{X_{n-3}} e^{\log \phi_{n-3} + [\log \mu_{n-2} + E_{n-2} + \epsilon_{n-1} + \epsilon_n]} - \log \mu_{n-3}$$

yielding

$$\log \lambda_{n-3} - \log \mu_{n-3} \leq E_{n-3} + \epsilon_{n-2} + \epsilon_{n-1} + \epsilon_n$$

and so on. Clearly the emerging expression for bucket B_c is

$$\log \lambda_c - \log \mu_c \leq E_c + \epsilon_{c+1} + \epsilon_{c+2} + ... \tag{8}$$

or,

$$\log \lambda_c - \log \mu_c \leq E_c + \sum_{k=0}^{n-c-1} \epsilon_{c+1+k} \tag{9}$$

The general transition from $n-i$ to $n-i-1$ can be easily followed to complete the inductive proof. Assuming that we control the derivation of μ_c for each B_c to ensure that $E_c = \log \mu^*_c - \log \mu_c \leq \epsilon_c$ and substituting in the expression we get from Eq. (9) that

$$\log \lambda_c - \log \mu_c \leq \epsilon_c + \sum_{k=0}^{n-c-1} \epsilon_{c+1+k} \leq (n-c+1) * \epsilon \tag{10}$$

0.3 MISCALLANEOUS EXPERIMENTS ON ANALYZING ERROR

Calculating ϵ from Theorem 1 is hard because it involves computing the local bucket error E over all configurations in the scope of the bucket. Therefore, we calculate the maximum over a sampled test set (lines 18-19 of Algorithm 3) as $\hat{\epsilon}$. Additionally, we also calculate the average local bucket error, $\hat{\epsilon}^{avg}$ over the same test set.

To bound the global error of the approximated partition function from Eq. 3, we sum over all the estimated bucket error bounds, $\hat{\epsilon}$), Clearly, the bound is very lose. We therefore also use the average local bucket error, $\hat{\epsilon}^{avg}$ to give us some additional information on the global error empirically:

$$\hat{E}_1 \leq \sum_{k=0}^{n-1} \hat{\epsilon}^{avg}_{1+k} \tag{11}$$

Relationship between local and global errors empirically.

Figure 1a) depicts the empirical global errors against the local error bound for 4 grid instances over 2 sample configurations={60k,120k}. Specifically, the local error bound shown is the maximum over the estimated local bucket errors ($\hat{\epsilon}^{avg}$ from Eq. 11) across all buckets and the (empirical) global error is the error in the partition function estimate. As expected, we see a somewhat linear relationship between the global error and the local error bound. We also see that higher samples drive the local and global errors towards the lower-left of the plot and vice-versa.

Impact of sample size on error bounds. Figure 1 also depicts the impact of sample size on the estimated local and global error bounds (Eq. 11). Specifically, the local error bound shown is the maximum over the estimated local bucket errors ($\hat{\epsilon}^{avg}$ from Eq. 11) across all buckets. As expected, we see that increasing the training sample size makes the two bounds tighter. For the 4 grid instances (f10, f5, f2, f15), we also observed that the empirical global error in the partition function estimate for for the two sample configurations {(37.21, 7.94), (18.8, 5.9), (10.28, 3.05), (41.3, 27.5)} is in proportion to the global error bound from Fig. 1b).

	N_{avg} = 60k, h=w								N_{avg} = 150k, h=w							
	Statistics on local bucket errors				Statistics on global errors				Statistics on local bucket errors				Statistics on global errors			
Id	test w.m.s.e		local bucket errors		estimated bounds		empirical errors		test w.m.s.e		local bucket errors		estimated bounds		empirical errors	
	avg	max	avg	max	avg	max	avg	max	avg	max	avg	max	avg	max	avg	max
1	2.19E-04	0.06	1.67	3.81	251.7	1664	24.01	37.21	1.63E-05	0.006	0.29	3.71	71.75	1380	8.11	13.62
2	1.74E-04	0.053	0.784	2.44	115.6	796	21.15	31.32	1.40E-05	0.0068	0.131	1.45	34.34	680	6.4	12.3
3	1.06E-04	0.053	0.336	0.661	34.98	264.91	5.05	7.88	7.46E-06	0.004	0.045	0.46	10.28	228.59	4.3	5.95
4	1.94E-04	0.057	1.988	10.97	359.38	2549.28	22.43	27.68	1.85E-05	0.009	0.382	4.665	106.75	2166	31.03	57.46

Figure 1: Statistics of Local & bucket errors compared with global error over 5 runs for 4 grid-hard instances having w=55 with $i-$bound=20, where h=w, # buckets trained, $\#NB = 308$ for two different scales of smaples sizes. *test wmse* is the w.m.s.e of the learned NN over the test set; *local bucket error* is the average L1 error for $log\lambda$ approximations over all buckets; *estimated bounds* is the bound obtained in eq 3; *empirical error* is the average global error over 5 runs.

References

Peter L. Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(63):1–17, 2019. URL `http://jmlr.org/papers/v20/17-612.html`.