Supplementary Material For
Approximate Algorithms for Learning Bayesian Neural Networks

1 Rectified Linear Units

In the forward pass, we need to compute $\mathbb{E}\left[z_{il}\right]$ and $\mathbb{E}\left[z_{il}^2\right]$. Here we derive the expression in Equation 8 of the main text.

We know that $z_{il} = max(0, u_{il})$. The expectation can be computed as follows,

$$\begin{aligned}\mathbb{E}\left[z_{il}\right] &= \int_{-\infty}^{+\infty} \max(0, u_{il}) \mathcal{N}(u_{il} \mid \mu_{il}, \tau_{il}) du_{il} \\ &= \int_{0}^{\infty} u_{il} \mathcal{N}(u_{il} \mid \mu_{il}, \tau_{il}) du_{il}\end{aligned} \tag{1}$$

Dropping the subscripts and substituting $m = \frac{u-\mu}{\tau^{1/2}}$ we have,

$$\begin{aligned}&\mathbb{E}\left[z_{il}\right] = \int_{\frac{-\mu}{\sqrt{\tau}}}^{\infty} (\mu + \tau^{1/2} m) \frac{exp(-m^2/2)}{\sqrt{2\pi}} dm \\ &= \mu \int_{\frac{-\mu}{\sqrt{\tau}}}^{\infty} \frac{exp(-m^2/2)}{\sqrt{2\pi}} dm + \sqrt{\tau} \int_{\frac{-\mu}{\sqrt{\tau}}}^{\infty} m \frac{exp(-m^2/2)}{\sqrt{2\pi}} dm \\ &= \mu \Phi\left(\frac{\mu}{\sqrt{\tau}}\right) + \tau \mathcal{N}(\mu \mid 0, \tau)\end{aligned} \tag{2}$$

Next, we show how to compute the second moment.

$$\mathbb{E}\left[z_{il}^2\right] = \int_{0}^{\infty} u_{il}^2 \mathcal{N}(u_{il} \mid \mu_{il}, \tau_{il}) du_{il} \tag{3}$$

Again dropping the subscripts and substituting $m = \frac{u-\mu}{\tau^{1/2}}$ we have,

$$\begin{aligned}&\mathbb{E}\left[z_{il}^2\right] = \int_{\frac{-\mu}{\sqrt{\tau}}}^{\infty} (\mu + \tau^{1/2} m)^2 \frac{exp(-m^2/2)}{\sqrt{2\pi}} dm \\ &= \mu^2 \Phi\left(\frac{\mu}{\sqrt{\tau}}\right) + \frac{2\mu\sqrt{\tau}}{\sqrt{2\pi}} exp(-\mu^2/2\tau) + \tau \int_{\frac{-\mu}{\sqrt{\tau}}}^{\infty} m^2 \frac{e^{-m^2/2}}{\sqrt{2\pi}} dm \\ &= \mu^2 \Phi\left(\frac{\mu}{\sqrt{\tau}}\right) + \frac{2\mu\sqrt{\tau}}{\sqrt{2\pi}} exp(-\mu^2/2\tau) - \\ &\frac{\sqrt{\tau}\mu}{\sqrt{2\pi}} exp(-\mu^2/2\tau) + \tau \Phi\left(\frac{\mu}{\sqrt{\tau}}\right)\end{aligned} \tag{4}$$

Rearranging terms we have,

$$\mathbb{E}\left[z_{il}^2\right] = (\mu^2 + \tau) \Phi\left(\frac{\mu}{\sqrt{\tau}}\right) + \mu\tau \mathcal{N}(\mu \mid 0, \tau) \tag{5}$$

2 Multiclass posterior predictive distribution

The posterior predictive distribution for a new feature x_* can be calculated through a Monte Carlo approximation.

$$\begin{aligned}p(y_* \mid x_*, \mathcal{D}) &= \int p(y_* \mid x_*, \mathcal{W}) p(\mathcal{W}, \lambda \mid \mathbf{y}, \mathbf{x}) d\mathcal{W} d\lambda \\ &\approx \int p(y_* \mid x_*, \mathcal{W}) q(\mathcal{W}, \lambda \mid \mathbf{y}, \mathbf{x}) d\mathcal{W} d\lambda \\ &= \int p(y_* \mid x_*, \mathcal{W}) q(\mathcal{W} \mid \vartheta) d\mathcal{W} \\ &\approx \int \sigma(\mathbf{z}_L) \mathcal{N}(\mathbf{z}_L \mid \nu_L, \Psi_L) d\mathbf{z}_L \\ &\approx \frac{1}{S} \sum_s \mathbf{z}_L^s, \quad \mathbf{z}_L^s \sim \mathcal{N}(\mathbf{z}_L \mid \nu_L, \Psi_L)\end{aligned} \tag{6}$$

Our experiments used $S = 100$ samples.

3 Continuous regression and Binary classification experiments

Descriptions of datasets

We used ten UCI regression datasets for comparing PBP against rectified linear EBP. Table 1 summarizes the characteristics of the different datasets. The order of the presented datasets correspond to the labeling 1 through 10 used in the main paper For binary classification, we used text classification datasets summarized in Table 2.

Test log likelihoods

In Figure 1 we present the per dataset test log-likelihoods achieved by PBP and EBP on continuous regression and binary classification. Notice that on a large majority PBP achieves higher predictive log-likelihoods.

Multi layer experiments

We also compared the performance of EBP and PBP on multilayer architectures. Table 3 we summarize results from 2 and 5 layer networks on regression datasets.

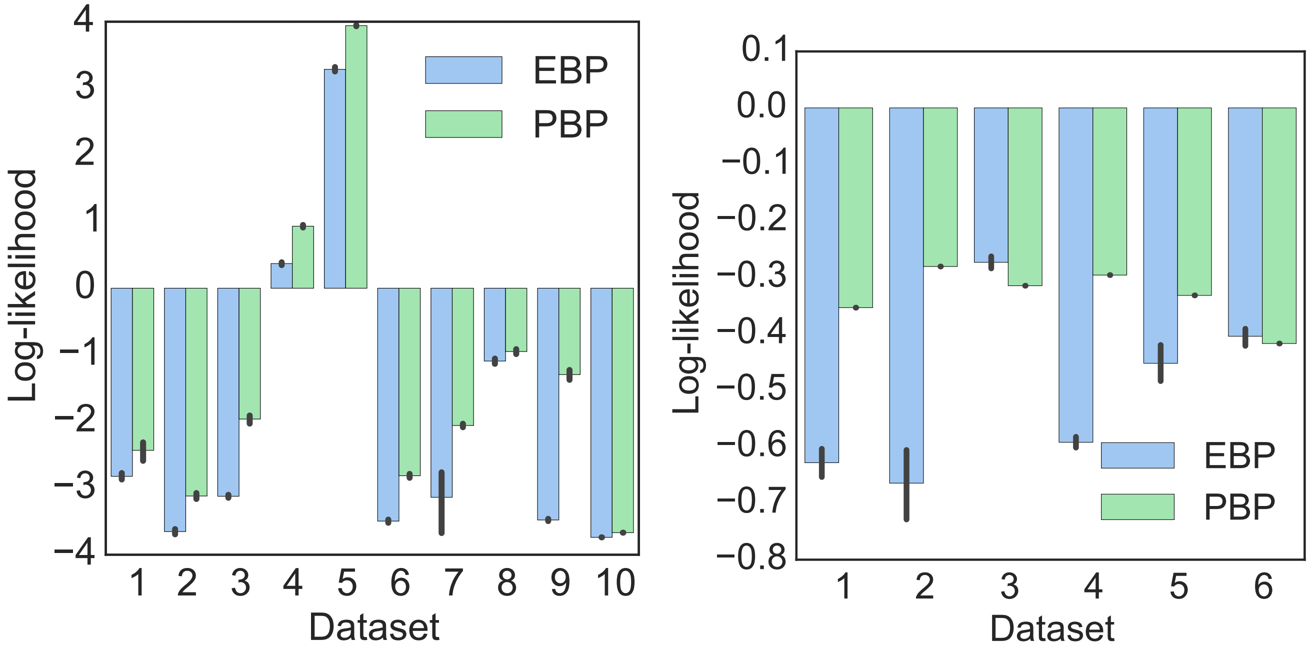

Figure 1: Test log likelihoods on regression (left) and classification datasets(right)

	Dataset	RMSE EBP 2layer	RMSE PBP 2layer	RMSE EBP 5layer	RMSE PBP 5layer
1	Boston	3.14 ± 0.93	$\mathbf{2.79 \pm 0.16}$	9.33 ± 1.0	3.08 ± 0.15
2	Concrete	5.30 ± 0.77	$\mathbf{5.24 \pm 0.11}$	6.33 ± 0.91	5.96 ± 0.16
3	Energy Efficiency	1.38 ± 0.17	$\mathbf{0.90 \pm 0.04}$	3.54 ± 3.03	1.18 ± 0.06
4	Kin8nm	0.07 ± 0.22	$\mathbf{0.07 \pm 0.00}$	0.18 ± 0.09	0.08 ± 0.00
5	Naval Propulsion	0.007 ± 0.00	$\mathbf{0.003 \pm 0.00}$	0.007 ± 0.00	0.004 ± 0.00
6	Power Plant	4.21 ± 0.23	$\mathbf{4.03 \pm 0.03}$	4.56 ± 0.25	4.08 ± 0.04
7	Protein Structure	$\mathbf{2.14 \pm 0.17}$	4.25 ± 0.02	2.04 ± 0.15	3.97 ± 0.04
8	Wine	0.71 ± 0.06	$\mathbf{0.64 \pm 0.00}$	0.82 ± 0.04	0.64 ± 0.01
9	Yacht	1.14 ± 0.45	$\mathbf{0.85 \pm 0.05}$	5.58 ± 5.77	1.71 ± 0.23
10	Year Prediction	9.21	8.21	NA	8.93

Table 3: RMSE test error rates for EBP and PBP using 2 layer and 5 layer architectures.

4 Multiclass Experiments

Here we present additional results for the multiclass experiments.

Log bound vs Stochastic approximation

We evaluated the differences between log bound and stochastic approximations by measuring their performance on three multi class datasets – MNIST, UCI HAR, a six class human activity recognition dataset and Sensorless Drive Diagnosis Data Set, a 11 class dataset for detecting malfunctioning components. On each dataset we trained a network with 2 hidden layers of 400 units each and trained for a 100 epochs using both PBP and EBP. Figure 2 summarizes the performance of the two approximations, averaged over all datasets and the two algorithms (EBP and PBP). This clearly demonstrates the superior performance of the stochastic approximation. For this experiment, we used 100 samples, but a similar trend holds even with a single sample.

Stochastic approximation quality

Figure 3 displays the variance of the training log likelihood for 1, 10 and 100 sample stochastic approximations, for datasets containing 10 and 100 classes. We see that the variance decreases with increasing number of samples.These results are for PBP, EBP performs similarly.

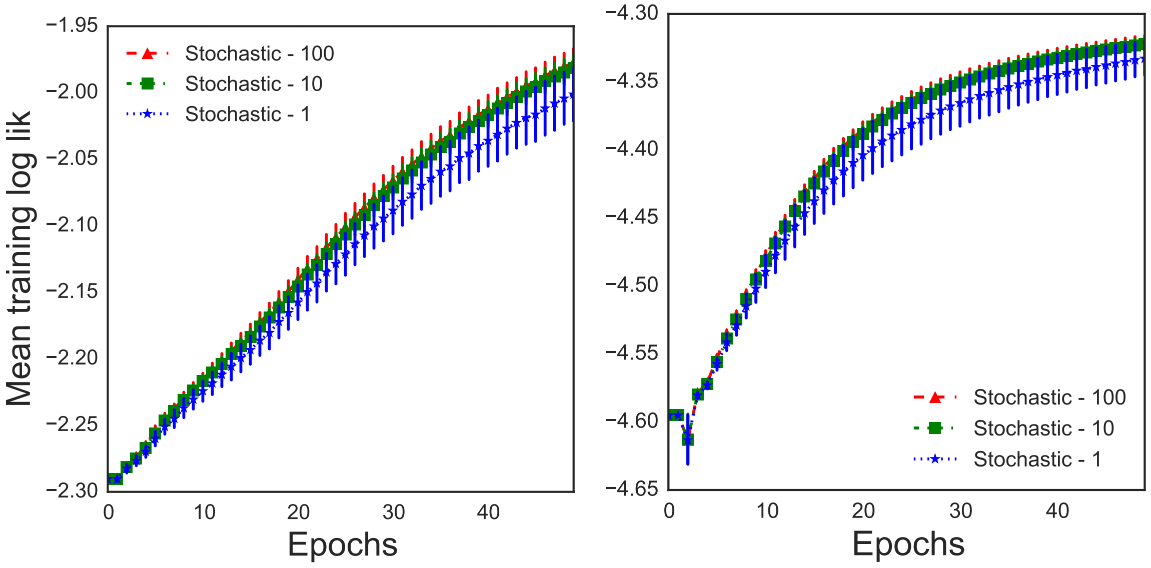

Figure 3: Training log likelihood variance on synthetic data, with 10 (left) and 100 (right) classes.

	Dataset	N	d
1	Boston	506	13
2	Concrete Compression Strength	1030	8
3	Energy Efficiency	768	8
4	Kin8nm	8192	8
5	Naval Propulsion	11,934	16
6	Combined Cycle Power Plant	9568	4
7	Protein Structure	9568	4
8	Wine Quality Red	1599	11
9	Yacht Hydrodynamics	308	6
10	Year Prediction MSD	515,345	90

Table 1: Continuous regression datasets

	Dataset	N	d
1	20News group comp vs HW	1943	29409
2	20News group elec vs med	1971	38699
3	Spam or ham d0	2500	26580
4	Spam or ham d1	2500	27523
5	Reuters news I8	2000	12167
6	Reuters news I6	2000	11463

Table 2: Binary classification datasets

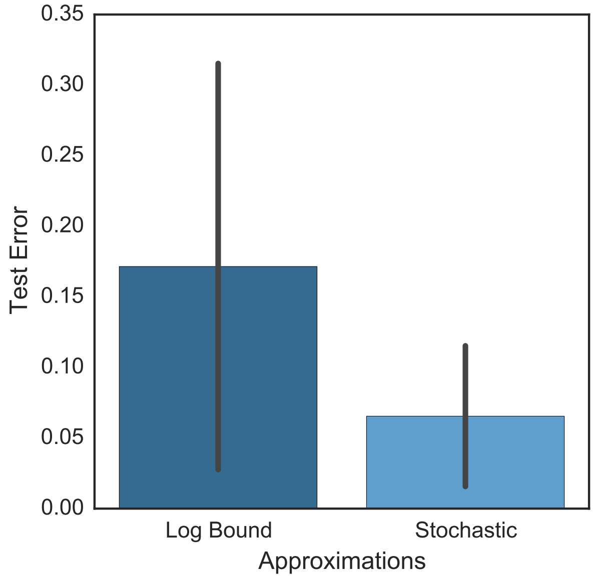

Figure 2: Performance of Log bound vs Stochastic approximations, averaged over algorithms (EBP and PBP) and datasets.