Probabilistic Backpropagation for Scalable Learning of Bayesian Neural Networks
José Miguel Hernández-Lobato\(^1\) and Ryan P. Adams\(^1\) 
\(^1\)Harvard University

1. Motivation

Multilayer neural networks are state-of-the-art techniques, but... 
- Require tuning of hyper-parameters. 
- Are affected by overfitting problems. 
- Lack estimates of uncertainty in their predictions.

The Bayesian approach can solve these problems but existing methods lack scalability, until now...

2. Probabilistic Multilayer Neural Networks

- **ReLU activations** for the hidden units: \( a(x) = \max(x, 0) \).
- The likelihood: \( p(y|W, X, \gamma) = \prod_{n=1}^{N} \mathcal{N}(y_n|z_n(W), \gamma^{-1}) = f_n \).
- The priors: \( p(W|\lambda) = \prod_{l=1}^{L} \prod_{i=1}^{V_l} \prod_{j=1}^{V_l-1} \mathcal{N}(w_{ij}|0, \lambda^{-1}) = g_{B} \).

The posterior approximation is \( q(W, \gamma, \lambda) = \prod_{l=1}^{L} \prod_{i=1}^{V_l} \prod_{j=1}^{V_l-1} \mathcal{N}(w_{ij}|m_{ij}, v_{ij}) \Gamma \mathcal{m}(\alpha, \beta) \Gamma \mathcal{m}(\alpha, \beta) \).

3. Probabilistic Backpropagation

After seeing the \( n \)-th data point, Bayes rule updates our beliefs \( q(w) \) as

\[
p(w) = Z^{-1} q(w|z_n(W), \gamma^{-1})
\]

where \( Z \) is the normalization constant.

PBP uses \( q(w) = \mathcal{N}(w|m, v) \) and approximates \( p(w) \) with \( \mathcal{N}(w|m^{\text{new}}, v^{\text{new}}) \)

We match moments between \( p(w) \) and \( q(w) \).

4. Forward Pass

Propagate distributions through the network and approximate them with Gaussians by moment matching.

5. Results on Toy Dataset

40 training epochs. 
100 hidden units. 
VI uses two stochastic approximations to the ELBO.

BP and VI tuned with Bayesian optimization (www.whetlab.com).

6. Results Predictive Performance and Running Time

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Avg. Test RMSE and Std. Errors</th>
<th>Avg. Test LL and Std. Errors</th>
<th>Running Time in Seconds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boston Housing</td>
<td>4.320±0.20</td>
<td>3.644±0.08</td>
<td>1035</td>
</tr>
<tr>
<td>Concrete Compression Strength</td>
<td>7.128±0.00</td>
<td>6.307±0.04</td>
<td>1005</td>
</tr>
<tr>
<td>Energy Efficiency</td>
<td>6.509±0.00</td>
<td>5.866±0.02</td>
<td>2895</td>
</tr>
<tr>
<td>Kiva</td>
<td>8.192±0.00</td>
<td>7.690±0.02</td>
<td>2585</td>
</tr>
<tr>
<td>Natural Propulsion</td>
<td>11.324±0.00</td>
<td>10.995±0.02</td>
<td>2985</td>
</tr>
<tr>
<td>Combined Cycle Power Plant</td>
<td>9.584±0.00</td>
<td>9.127±0.02</td>
<td>2985</td>
</tr>
<tr>
<td>Protein Structure</td>
<td>45.730±0.00</td>
<td>43.423±0.02</td>
<td>7995</td>
</tr>
<tr>
<td>Wine Quality Red</td>
<td>15.991±0.00</td>
<td>14.645±0.00</td>
<td>1195</td>
</tr>
<tr>
<td>Yacht Hydrodynamics</td>
<td>308±0.00</td>
<td>287±0.00</td>
<td>9455</td>
</tr>
<tr>
<td>Year Prediction MSD</td>
<td>515.345±0.00</td>
<td>503.425±0.00</td>
<td>142.007</td>
</tr>
</tbody>
</table>

7. Results with More than One Hidden Layer

8. Results Active Learning

9. Summary

- PBP is a state-of-the-art method for scalable inference in NNs. 
- PBP is very similar to traditional backpropagation. 
- PBP often outperforms backpropagation at a lower cost. 
- Very fast C code available at https://github.com/HIPS