Bayesian Deep Learning

Extending Probabilistic Backpropagation and Transfer Learning

The University of Texas at Austin
Department of Statistics and Data Sciences
College of Natural Sciences

Evan Ott

Advisor: Sinead Williamson

May 3, 2018

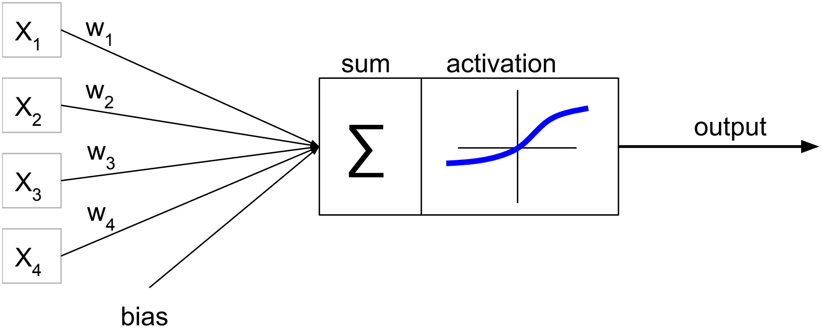
Overview

- Neural Networks
- Bayesian Neural Networks
  - Laplace Approximation
  - Variational Inference
  - Assumed Density Filtering
  - Probabilistic Backpropagation
- Preliminary Work
- Planned Contributions

Perceptron

Basic unit of neural network, described by:

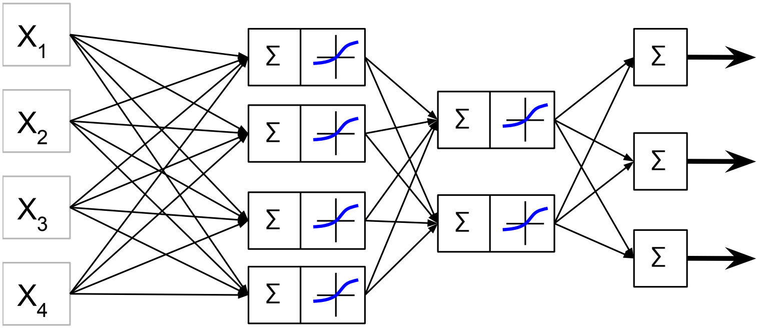
$$f(\mathbf{x}) = \sigma(\mathbf{x}^\top \omega)$$



Frank Rosenblatt. *The Perceptron, a Perceiving and Recognizing Automaton*. Tech. rep. Cornell Aeronautical Laboratory, 1957.

Multi-layer Perceptron

Many perceptrons arranged in layers



Backpropagation (BP)

- BP minimizes a cost function:

  Regression $\|Y - \mathtt{NN}(\mathbf{x}; \mathcal{W})\|_2^2$

  Classification $-\sum_{k=1}^{K} Y_k \log(\mathtt{NN}_k(\mathbf{x}; \mathcal{W}))$

- Optional regularization:

  L1 $\lambda \sum_{l,i,j} |W_{lij}|$

  L2 $\lambda \sum_{l,i,j} W_{lij}^2$

- Cleverly applies chain rule of derivatives

$$Z = e^{-Y}, \qquad Y = \frac{1}{1 + e^{-X}}$$

$$\frac{dZ}{dY} = -Z, \qquad \frac{dZ}{dX} = \frac{dZ}{dY}\frac{dY}{dX} = -ZY(1 - Y)$$

Deep Neural Networks

Advantages:

- Flexible, fast to train (e.g., SGD with backpropagation)
- Can achieve high accuracy/precision/recall
  - Identifying objects in images (Szegedy et al. 2015)
  - Melanoma detection from images (Esteva et al. 2017)
  - Tuberculosis detection from chest x-rays (Lakhani and Sundaram 2017)

Drawbacks:

- Typically, only provides point estimates
- Tendency for overfitting
- Unclear choice of structure

Christian Szegedy et al. Going Deeper with Convolutions. In: *Computer Vision and Pattern Recognition*. 2015.

Andre Esteva et al. Dermatologist-level Classification of Skin Cancer with Deep Neural Networks. In: *Nature* 542.7639 (2017), p. 115.

Paras Lakhani and Baskaran Sundaram. Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by using Convolutional Neural Networks. In: *Radiology* 284.2 (2017), pp. 574–582.

From Backpropagation to Bayesian Neural Networks

Introduce joint probability model:

- Likelihood: $p(Y|\theta)$

  Regression $p(Y|\mathcal{W}, \sigma^2, \mathbf{x}) = N(Y; \mathtt{NN}(\mathbf{x}; \mathcal{W}), \sigma^2)$
  Classification $p(Y|\mathcal{W}, \mathbf{x}) = \mathrm{Categorical}(Y; \mathtt{NN}(\mathbf{x}; \mathcal{W}))$

- Prior distribution: $p(\theta)$

  Independent $w_{lij} \overset{\text{iid}}{\sim} N(0, \tau)$
  Correlated $w_l \sim MN(M, R, C)$, see (Louizos and Welling 2016)

- Posterior distribution:

$$p(\theta|Y) = \frac{p(Y|\theta)p(\theta)}{\int p(Y|\theta)p(\theta)d\theta}$$

Christos Louizos and Max Welling. Structured and Efficient Variational Deep Learning with Matrix Gaussian Posteriors. In: *International Conference on Machine Learning*. 2016, pp. 1708–1716.

Bayesian Neural Networks (BNNs)

The problem:

- Posterior (or posterior predictive, etc.) is intractable
- MCMC possible for small networks (Neal 1993)

Methods used for BNN inference:

- **Assumed density filtering**
- Dropout as deep GP (Gal and Ghahramani 2016)
- Expectation propagation (Soudry et al. 2014)
- **Laplace approximation**
- **Variational inference**

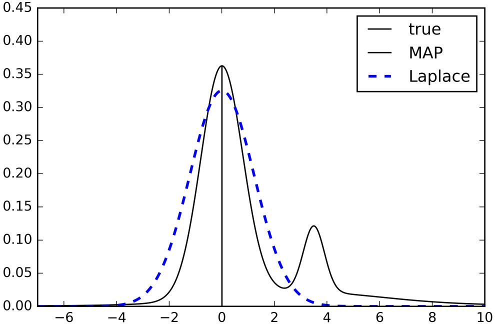
Radford M Neal. Bayesian Learning via Stochastic Dynamics. In: *Advances in Neural Information Processing Systems*. 1993, pp. 475–482.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In: *International Conference on Machine Learning*. 2016, pp. 1050–1059.

Daniel Soudry et al. Expectation Backpropagation: Parameter-free Training of Multilayer Neural Networks with Continuous or Discrete Weights. In: *Advances in Neural Information Processing Systems*. 2014, pp. 963–971.

Laplace Approximation

- Approximate posterior introduced by (MacKay 1992)
- Identify MAP estimate by standard backpropagation
- Locally-quadratic approximation to form a Gaussian



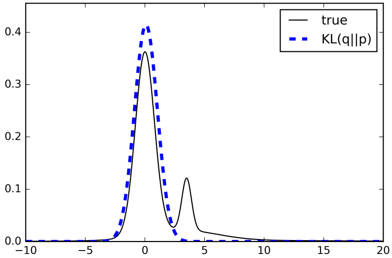
David JC MacKay. A Practical Bayesian Framework for Backpropagation Networks. In: *Neural Computation* 4.3 (1992), pp. 448–472.

Variational Inference

- Posit variational family $\mathcal{Q}$ to approximate $p(W|\mathcal{D})$
- Identify $q(W) \in \mathcal{Q}$ that minimizes

$$KL(q(W) \| p(W|\mathcal{D})) = \int_{\mathcal{W}} q(W) \log \left(\frac{q(W)}{p(W|\mathcal{D})} \right) dW$$

- Applied to BNNs by (Graves 2011) with MCMC likelihood



Alex Graves. Practical Variational Inference for Neural Networks. In: *Advances in Neural Information Processing Systems*. 2011, pp. 2348–2356.

Assumed Density Filtering

Approximate Bayesian approach to online learning (Opper and Winther 1998)

- Approximate posterior $q(\theta|\gamma_t)$ at iteration t with parameters γ_t
  - For example, if q is Gaussian, $\gamma_t = (\mu_t, \sigma_t^2)$
- Given new data y_{t+1}:

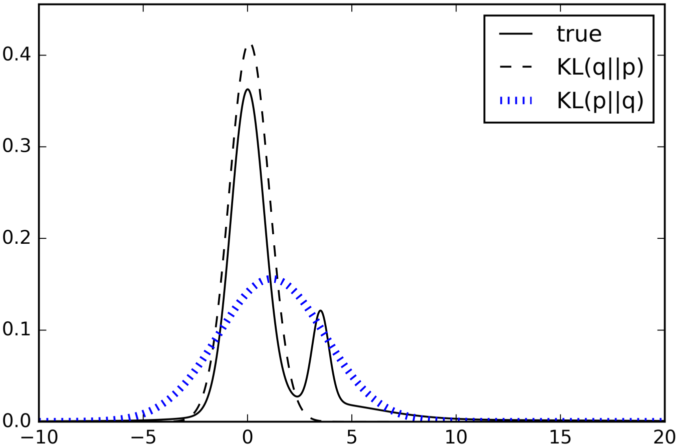
  Update Update "exact" posterior:

$$p(\theta|y_{t+1}, \gamma_t) = \frac{p(y_{t+1}|\theta)q(\theta|\gamma_t)}{\int p(y_{t+1}|\theta)q(\theta|\gamma_t)d\theta}$$

  Projection $\gamma_{t+1} := \arg\min_{\gamma} D\left(p(\cdot|y_{t+1}, \gamma_t) \parallel q(\cdot|\gamma)\right)$

Manfred Opper and Ole Winther. A Bayesian Approach to On-line Learning. In: *On-line Learning in Neural Networks, ed. D. Saad* (1998), pp. 363–378.

Assumed Density Filtering



Probabilistic Backpropagation (PBP)

ADF algorithm for BNNs (Hernández-Lobato and Adams 2015)

- Normal prior on all weights:

$$w_{lij} \overset{\text{iid}}{\sim} N(0, \tau)$$

- Independent normal approximate posterior on each weight:

$$q(w_{lij}) = N(w_{lij} | m_{lij}, v_{lij})$$

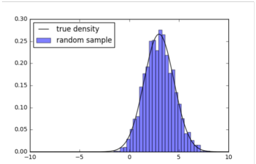
- Regression likelihood: $Y | \mathbf{x}, \mathcal{W} \sim N(\mathtt{NN}(\mathbf{x}; \mathcal{W}), \gamma^{-1})$
- Sequential closed-form approximation for normalization constant, posterior predictive

José Miguel Hernández-Lobato and Ryan Adams. Probabilistic Backpropagation for Scalable Learning of Bayesian Neural Networks. In: *International Conference on Machine Learning*. 2015, pp. 1861–1869.

Probabilistic Backpropagation (PBP)

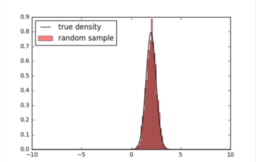
Closed-form approximations match moments to a Gaussian

Illustration with simulated data:



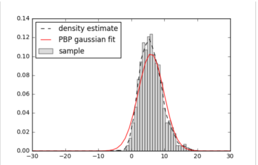
X~N(3, 2.25)

X

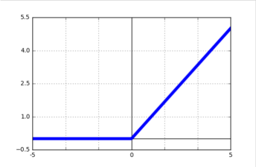


W~N(2, 0.25)

=

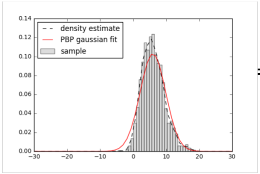


p(XW)



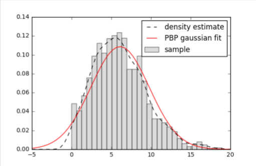
f(z)=max(0, z)

O



p(XW)

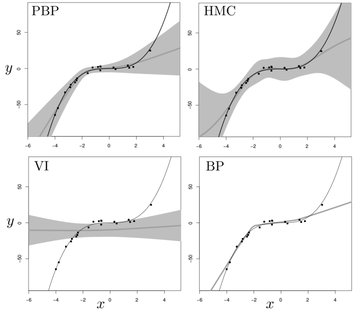
=



p(f(XW))

PBP Properties

- Trains like standard MLP, so it's fast
- Only given for regression with ReLU activation
- Extended by (Ghosh et al. 2016) to binary classification (probit) and multiclass via MCMC step



Soumya Ghosh et al. Assumed Density Filtering Methods for Learning Bayesian Neural Networks. In: *AAAI Conference on Artificial Intelligence*. 2016, pp. 1589–1595.

Hernández-Lobato and Adams 2015, Figure 1.

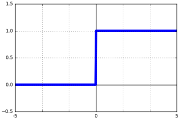
Comparison

	MCMC	VI	PBP
Object	w	$q(w)$	$q(w)$
Strategy	Simulation	Optimization	Optimization
Integrals	Sample MCMC chain	Sample $q(w)$	Closed-form approximations
Scale	?	Y	Y
Speed	N	?	Y
Papers	(Neal 1993)	(Graves 2011)	(Hernández-Lobato and Adams 2015)

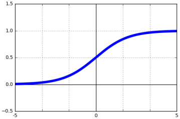
Preliminary Work

Exploring other activation functions

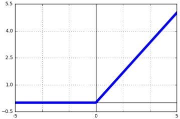
Step $\sigma(z) = \mathbb{1}\{z \geq 0\}$



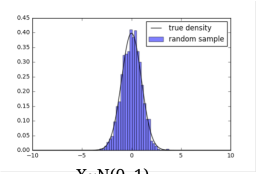
Sigmoid $\sigma(z) = 1/(1 + e^{-z})$



ReLU $\sigma(z) = \max(0, z)$

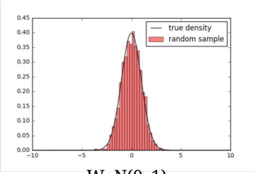


Preliminary Work



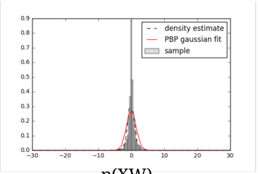
X~N(0, 1)

X

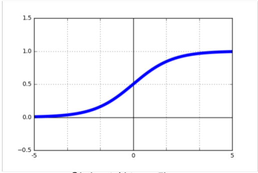


W~N(0, 1)

=

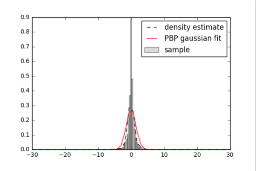


p(XW)



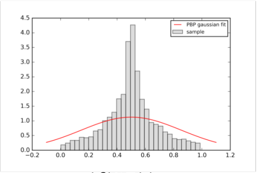
$f(z)=1/(1+e^{-z})$

O



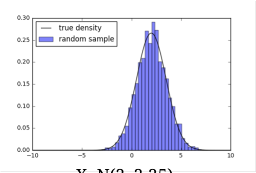
p(XW)

=



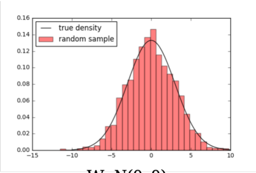
p(f(XW))

Preliminary Work



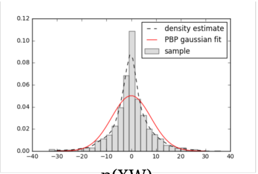
X~N(2, 2.25)

X

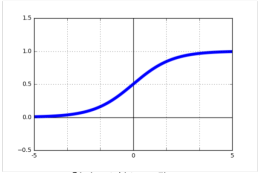


W~N(0, 9)

=

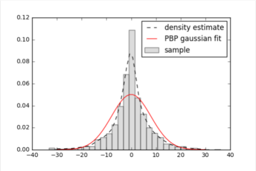


p(XW)



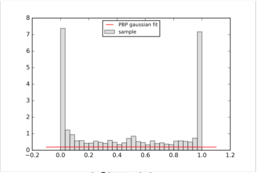
$f(z)=1/(1+e^{-z})$

O



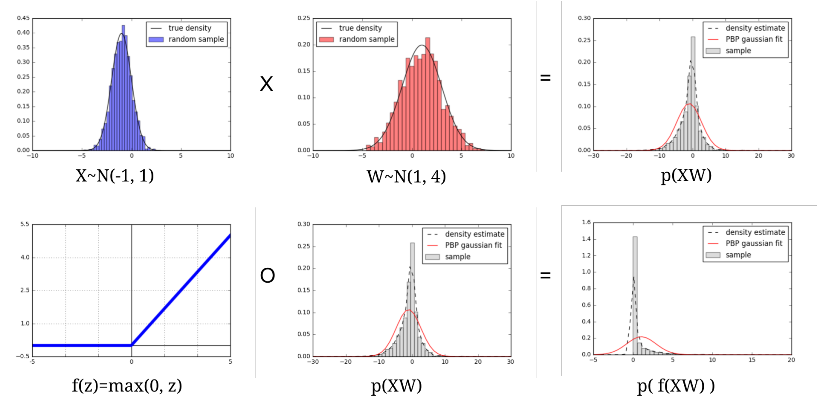
p(XW)

=



p(f(XW))

Current and Past Work



Led to questions about Gaussian approximation - replace with spike and slab?

$$q(w) = (1 - \pi)\delta_0(w) + \pi N(w; \mu, \sigma^2)$$

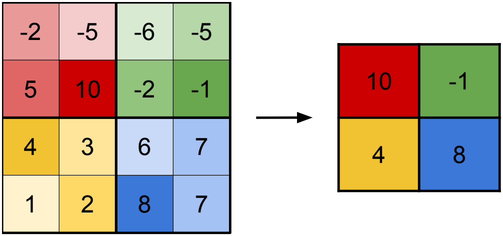
Planned Contributions

Classification Alternative to softmax without MCMC?

$$\hat{p}_i = \mathrm{Softmax}_i(\mathbf{x}) = \frac{e^{x_i}}{\sum_j e^{x_j}}$$

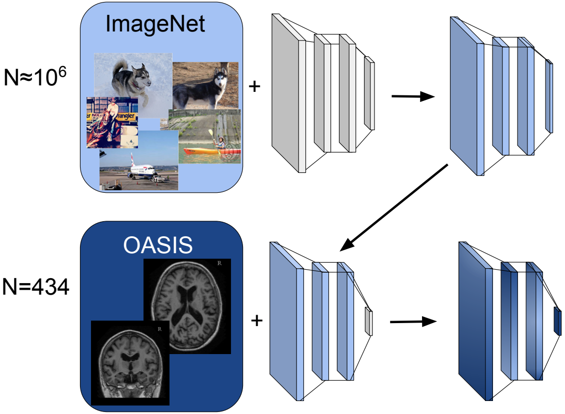
$$\tilde{p}_i = p(\mathtt{NN}(\mathbf{x}; \mathcal{W}) \in \mathcal{A}_i)$$

Pooling Need approximation for $p\left(\max\left(X_1, X_2, \ldots, X_k\right)\right)$



Planned Contributions

Bayesian version of transfer learning



Olga Russakovsky et al. ImageNet Large Scale Visual Recognition Challenge. In: *International Journal of Computer Vision* 115.3 (2015), pp. 211–252.

Daniel S Marcus et al. Open Access Series of Imaging Studies: Longitudinal MRI Data in Nondemented and Demented Older Adults. In: *Journal of Cognitive Neuroscience* 22.12 (2010), pp. 2677–2684.

Summary

- Deep neural networks in real-world problems
- Quantifying uncertainty critical for decision-making
- PBP as scalable, flexible BNN framework
- Need Bayesian analogues for deep learning practices

Thanks

Questions?

References I

- Esteva, Andre, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level Classification of Skin Cancer with Deep Neural Networks. In: *Nature* 542.7639 (2017), p. 115.
- Gal, Yarin and Zoubin Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In: *International Conference on Machine Learning*. 2016, pp. 1050–1059.
- Ghosh, Soumya, Francesco Maria Delle Fave, and Jonathan S Yedidia. Assumed Density Filtering Methods for Learning Bayesian Neural Networks. In: *AAAI Conference on Artificial Intelligence*. 2016, pp. 1589–1595.
- Graves, Alex. Practical Variational Inference for Neural Networks. In: *Advances in Neural Information Processing Systems*. 2011, pp. 2348–2356.

References II

- Hernández-Lobato, José Miguel and Ryan Adams. Probabilistic Backpropagation for Scalable Learning of Bayesian Neural Networks. In: *International Conference on Machine Learning*. 2015, pp. 1861–1869.
- Lakhani, Paras and Baskaran Sundaram. Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by using Convolutional Neural Networks. In: *Radiology* 284.2 (2017), pp. 574–582.
- Louizos, Christos and Max Welling. Structured and Efficient Variational Deep Learning with Matrix Gaussian Posteriors. In: *International Conference on Machine Learning*. 2016, pp. 1708–1716.
- MacKay, David JC. A Practical Bayesian Framework for Backpropagation Networks. In: *Neural Computation* 4.3 (1992), pp. 448–472.

References III

- Marcus, Daniel S, Anthony F Fotenos, John G Csernansky, John C Morris, and Randy L Buckner. Open Access Series of Imaging Studies: Longitudinal MRI Data in Nondemented and Demented Older Adults. In: *Journal of Cognitive Neuroscience* 22.12 (2010), pp. 2677–2684.
- Neal, Radford M. Bayesian Learning via Stochastic Dynamics. In: *Advances in Neural Information Processing Systems*. 1993, pp. 475–482.
- Opper, Manfred and Ole Winther. A Bayesian Approach to On-line Learning. In: *On-line Learning in Neural Networks, ed. D. Saad* (1998), pp. 363–378.
- Rosenblatt, Frank. *The Perceptron, a Perceiving and Recognizing Automaton*. Tech. rep. Cornell Aeronautical Laboratory, 1957.

References IV

- Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. In: *International Journal of Computer Vision* 115.3 (2015), pp. 211–252.
- Soudry, Daniel, Itay Hubara, and Ron Meir. Expectation Backpropagation: Parameter-free Training of Multilayer Neural Networks with Continuous or Discrete Weights. In: *Advances in Neural Information Processing Systems*. 2014, pp. 963–971.

References V

Szegedy, Christian, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going Deeper with Convolutions. In: *Computer Vision and Pattern Recognition*. 2015.

Hamiltonian Monte Carlo

- Let q be the parameters of our distribution $P(q)$
- Define potential energy $E(q)$ as $P(q) \propto \exp(-E(q))$
- Augment space to include momentum vector p, same dimension as q
- Define Hamiltonian $H(q, p) = E(q) + \frac{1}{2}\|p\|_2^2$
- Use Hamiltonian dynamics for equal-energy trajectories:

$$\frac{dq}{dt} = \frac{\partial H}{\partial p} = p \qquad \frac{dp}{dt} = -\frac{\partial H}{\partial q} = -\nabla E(q)$$

- Use log posterior
  $-\log P(q) = -\log f(X|q) - \log \pi(q) + \log p(X)$
- Find valid state, give it a kick, follow trajectory, move via Metropolis-Hastings.

Dropout as Variational Inference

(Gal and Ghahramani 2016)

- Stochastically set nodes in network to 0
- Connection to deep Gaussian process
- Really, dropout is a regularizer
- Matt Taddy and others: variational dropout provides poor variance estimates

Gal and Ghahramani 2016.