Weight Uncertainty in Neural Networks*

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, Daan Wierstra

Presentation by Blair Bilodeau[1] and Daniel Severo[1]

February 25, 2021

Presented to STA4273

[1]University of Toronto and Vector Institute

*C. Blundell et al. "Weight Uncertainty in Neural Networks". In: *Proceedings of the 32nd International Conference on Machine Learning*. 2015

Predictive Uncertainty

Observe: $X_{1:n} \sim \nu^{\otimes n}$ and $Y_{1:n} = \mathrm{Noise}\Big[f^*(X_{1:n})\Big]$.

Goal: Predict $Y \mid X$.

Solution: Approximate $\mathbb{E}[Y \mid X]$ with a neural network f_W.

Performance: Measured by $R(W) = \mathbb{E}(f_W(X) - Y)^2$.

Problem: Two sources of uncertainty are not captured.

Aleatoric — the amount of noise may affect $Y \mid X$ in complicated ways.

Epistemic — the model may be uncertain of f_W for rare X's.

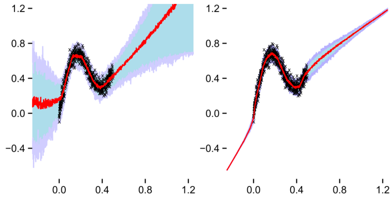

Figure 5. Regression of noisy data with interquatile ranges. Black crosses are training samples. Red lines are median predictions. Blue/purple region is interquartile range. Left: Bayes by Backprop neural network, Right: standard neural network.

Bayesian Prediction

Aleatoric — the amount of noise may affect $Y \mid X$ in complicated ways.
Epistemic — the model may be uncertain of f_W for rare X's.

To address aleatoric uncertainty, $f_W(X)$ parametrizes a *distribution* over Y.

If $Y \in \{0, 1\}$, $\hat{Y} \sim \mathrm{Bernoulli}(f_W(X))$.

If $Y \in \mathbb{R}$, $\hat{Y} \sim \mathrm{Normal}(f_W(X))$.

In general, we can define the *conditional likelihood* $p(Y \mid X, W)$.

Epistemic refers to uncertainty in f_W's estimate of the uncertainty parameter!

Epistemic uncertainty is quantified by the *posterior* $p(W \mid Y_{1:n}, X_{1:n})$.
Given a *prior* $\pi(W)$, we can write

$$p(W \mid Y_{1:n}, X_{1:n}) = \frac{p(Y_{1:n} \mid W, X_{1:n})\pi(W)}{p(Y_{1:n} \mid X_{1:n})}.$$

Let $\mathcal{D}_n = (X_i, Y_i)_{i \in [n]}$. For simplicity, we write

$$p(W \mid \mathcal{D}_n) = \frac{p(\mathcal{D}_n \mid W)\pi(W)}{p(\mathcal{D}_n)}.$$

Variational Inference

$$p(W \mid \mathcal{D}_n) = \frac{p(\mathcal{D}_n \mid W)\pi(W)}{p(\mathcal{D}_n)}.$$

Computing $p(\mathcal{D}_n)$ is intractable!

Instead, approximate $p(W \mid \mathcal{D}_n)$ with $q_\theta(W)$, a parametric density.

$$\begin{aligned}
\hat{\theta}_n &= \arg\min_\theta \mathrm{KL}\left(q_\theta(W) \| \, p(W \mid \mathcal{D}_n)\right) \\
&= \arg\max_\theta \mathbb{E}_{W\sim q_\theta}\left[\log \pi(W) - \log\left(\frac{q_\theta(W)}{p(\mathcal{D}_n \mid W)}\right)\right] \\
&= \arg\max_\theta \mathrm{ELBO}_n(q_\theta).
\end{aligned}$$

We want to solve this with gradient ascent:

$$\hat{\theta}_n^{(k+1)} = \hat{\theta}_n^{(k)} + \eta \nabla_\theta \mathrm{ELBO}_n(q_\theta).$$

Reparametrization

$$\hat{\theta}_n^{(k+1)} = \hat{\theta}_n^{(k)} + \eta \nabla_\theta \mathrm{ELBO}_n(q_\theta).$$

Computing $\nabla_\theta \mathrm{ELBO}_n(q_\theta)$ is intractable!

Suppose that $W \sim q_\theta \iff W = t(\theta, \varepsilon)$, where ε is independent standard noise.
E.g., $\varepsilon \sim \mathrm{Normal}(0,1)$ or $\varepsilon \sim \mathrm{Unif}([0,1])$.

Let $h_n(W, \theta) = \log \pi(W) - \log \left(\frac{q_\theta(W)}{p(\mathcal{D}_n \mid W)} \right)$.

$$\begin{aligned} \frac{\partial}{\partial \theta} \mathrm{ELBO}_n(q_\theta) &= \frac{\partial}{\partial \theta} \mathbb{E}_{W \sim q_\theta} h_n(W, \theta) \\ &= \frac{\partial}{\partial \theta} \mathbb{E}_\varepsilon h_n\Big(t(\theta, \varepsilon), \theta\Big) \\ &= \mathbb{E}_\varepsilon \left[\frac{\partial h_n(t, \theta)}{\partial t} \frac{\partial t(\theta, \varepsilon)}{\partial \theta} + \frac{\partial h_n(t, \theta)}{\partial \theta} \right]_{t = t(\theta, \varepsilon)}. \end{aligned}$$

$g_n(\theta, \varepsilon) = \left[\frac{\partial h_n(t,\theta)}{\partial t} \frac{\partial t(\theta,\varepsilon)}{\partial \theta} + \frac{\partial h_n(t,\theta)}{\partial \theta} \right]_{t = t(\theta, \varepsilon)}$ is an *unbiased gradient estimator*!

Gaussian Example

Define $\theta = (\mu, \sigma)$ so that $q_\theta = \mathrm{Normal}(\mu, \mathrm{diag}(\sigma))$.
Reparametrize with $\sigma = \log(1 + e^\rho)$.
Then $t(\theta, \varepsilon) = \mu + \log(1 + e^\rho)\varepsilon$ where $\varepsilon \sim \mathrm{Normal}(0, I)$...
...and $\partial t(\theta, \varepsilon)/\partial\theta = \varepsilon[1 + e^{-\rho}]^{-1}$.

Evaluating at $t = t(\theta, \varepsilon)$...

$$
\begin{aligned}
g_n(\theta, \varepsilon) = & \frac{\varepsilon}{1 + e^{-\rho}} \frac{\partial}{\partial t} \log \pi(t) && \text{(prior)} \\
& + \frac{\varepsilon}{1 + e^{-\rho}} \frac{\partial}{\partial t} \log p(\mathcal{D}_n \mid t) && \text{(likelihood)} \\
& + \frac{\varepsilon}{1 + e^{-\rho}} \frac{\partial}{\partial t} \log q_\theta(t) + \frac{\partial}{\partial \theta} \log q_\theta(t) && \text{(approximate posterior)}
\end{aligned}
$$

The prior derivative can be achieved by autograd – this is data independent.
The likelihood derivative is a composition of an easy derivative (say a Normal) and the usual derivative of the neural net output computed by backprop.
The posterior derivative is easy by the chosen form of q_θ.

Prediction Intervals

How should we use our uncertainty quantifications?

Frequentist: $Y \mid X \sim \mathrm{Dist}(f_W(X))$ for a learned (fixed) W.

Output 95% quantile range of $\mathrm{Dist}(f_W(X))$ for each new X.

Bayesian: $Y \mid X \sim \mathrm{Dist}(f_W(X))$ for $W \sim p(W \mid \mathcal{D}_n)$.

Maximum a posteriori: $\widehat{W} = \arg\max p(W \mid \mathcal{D}_n)$ and use frequentist interval.

Doesn't even require VI! Ignores epistemic uncertainty – just regularized MLE.

Model averaging: $p(Y \mid X) = \mathbb{E}_{W \sim q_\theta(W)}[p(Y \mid X, W)]$.

Paper proposes this, but it still ignores some epistemic uncertainty!

Monte Carlo: Sample $\widehat{W} \sim q_\theta(W)$ and $Y \mid X \sim \mathrm{Dist}(f_{\widehat{W}}(X))$.

Repeat many times to get empirical 95% quantile range for $Y \mid X$.

E.g., $p(W \sim \mathcal{D}_n) = \mathrm{Normal}(\mu_W, \sigma_W)$ and $f_W(X) = (\mu, \sigma)$.

Sample $\widehat{W} \sim \mathrm{Normal}(\mu_W, \sigma_W)$, $f_{\widehat{W}}(X) = (\hat{\mu}, \hat{\sigma})$, $Y \mid X \sim \mathrm{Normal}(\hat{\mu}, \hat{\sigma})$.

Importance Sampling: Sample $\widehat{W}_{1:M} \sim q_\theta(W)$ and $Y_{1:M} \mid X \sim \mathrm{Dist}(f_{\widehat{W}_{1:M}}(X))$.

Only keep Y_j with (unnormalized) probability $p(\widehat{W}_j, \mathcal{D}_n)/q_\theta(\widehat{W}_j)$.

Approximates sampling from $p(\widehat{W}_j \mid \mathcal{D}_n)$, but may have higher variance.

Summary

Goal: Quantify uncertainty of predictions on rare covariates in addition to inherent uncertainty due to systemic noise.
Solution: Treat the weights as random and find their posterior using VI.

In theory, this should work as long as the variational inference is accurate...
...but in practice, training is a nightmare!

"The BNN posterior distribution is complicated and high dimensional, and it's really hard to approximate it accurately with fully factorized Gaussians."

– Roger Grosse and Jimmy Ba's slides

There are many other tools to address these stability and accuracy issues...

minibatching, SGLD over GD, dropout, etc.

It remains open to *fully* account for both sources of uncertainty in predictions.