What my deep model doesn't know...

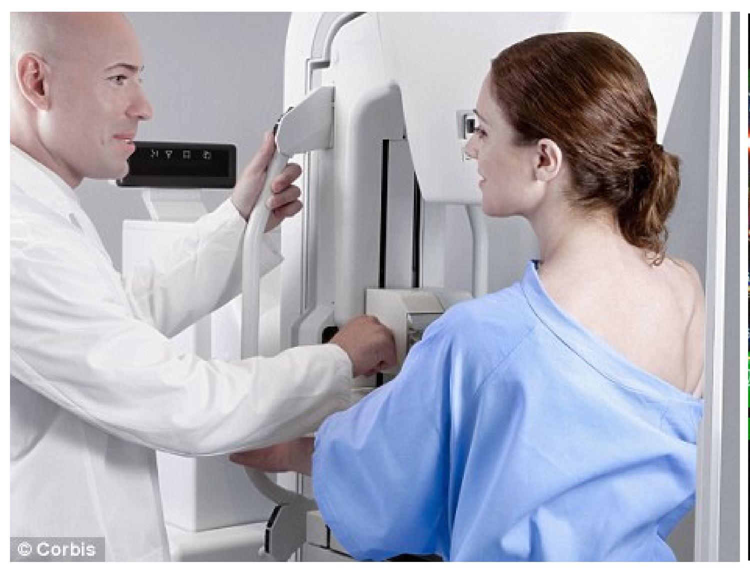
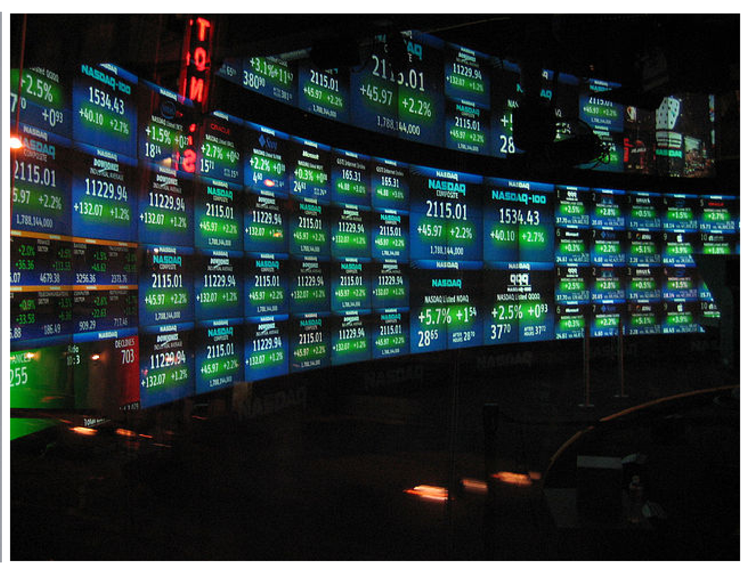
Principled and practical uncertainty estimates in deep learning without changing a thing.

Yarin Gal (`yg279@cam.ac.uk`), Zoubin Ghahramani (`zg201@cam.ac.uk`), University of Cambridge

UNIVERSITY OF CAMBRIDGE

Uncertainty is Everywhere.


Train a model to recognise dog breeds; someone will try to classify a cat.

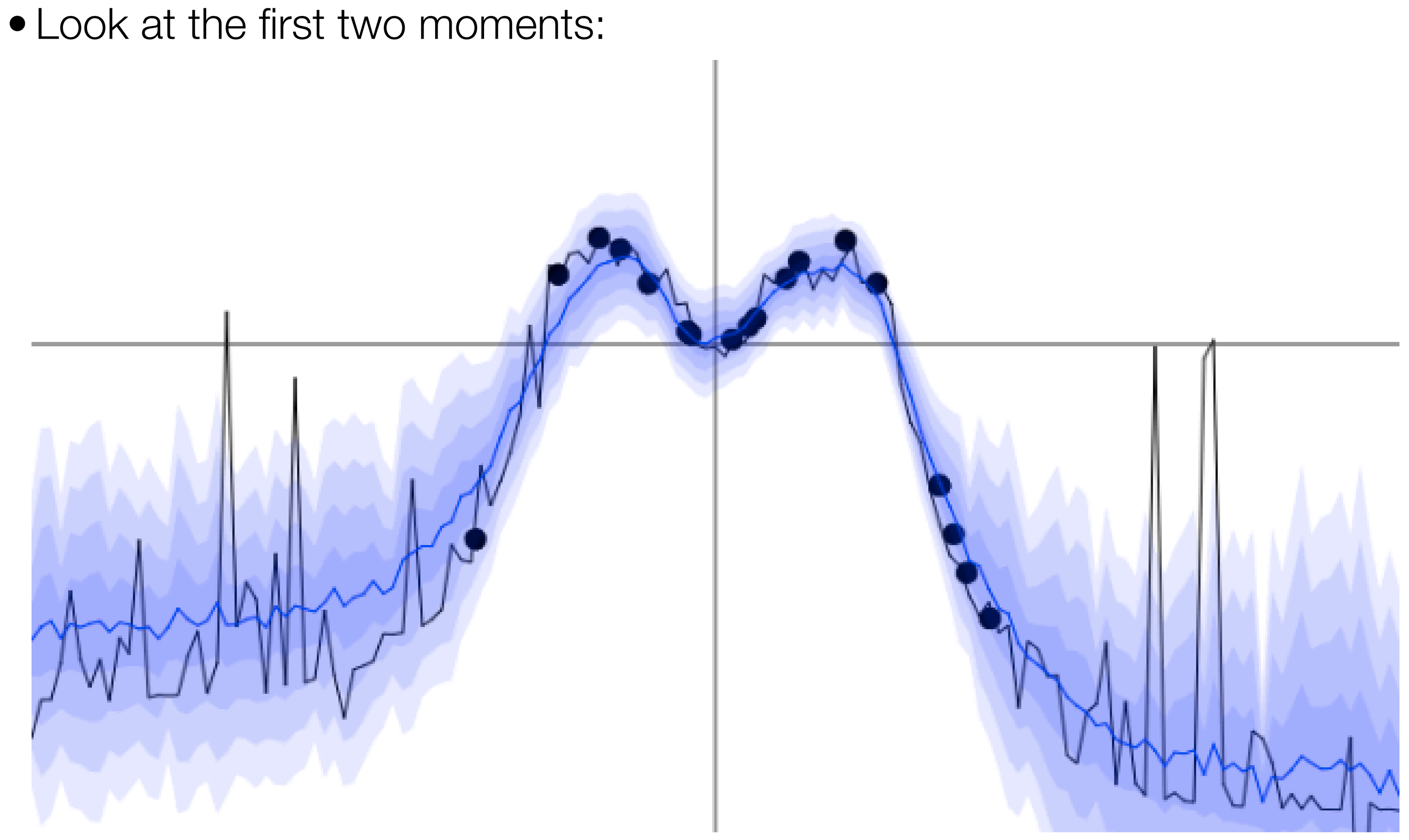
But also in decision making, life sciences, medicine, bioinformatics, self-driving cars, algo-trading...

For the practitioner:

- model diagnosis – *model should be certain about what it should know*
- use specialised models – *with simple and fast models for most data*
- critical systems – *pass data to a human to decide*

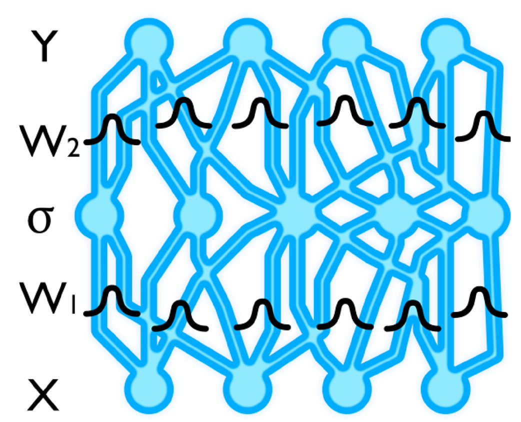
Uncertainty in Bayesian modelling

- Observed inputs $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N$ and outputs $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^N$
- Capture distribution believed to have generated outputs
- Look at the first two moments:



From Bayesian modelling to Dropout

- Place prior dist. $p(\mathbf{W})$ on weights, making these r.v.s



- Given dataset $\mathbf{X}, \mathbf{Y}$, the r.v. $\mathbf{W}$ has a posterior: $p(\mathbf{W}|\mathbf{X}, \mathbf{Y})$
- Which is difficult to evaluate...
- We can define a simple distribution $q_\theta(\cdot)$ and approximate

$$q_\theta(\mathbf{W}) \approx p(\mathbf{W}|\mathbf{X}, \mathbf{Y})$$

- Inference with

$$q_\theta(\mathbf{W}) := \mathbf{M} \cdot \text{diag(Bernoulli)}$$

and parameter $\mathbf{M}$

= Dropout training.

Practical Uncertainty Estimates

Using dropout we fit a distribution...

- Use first moment for predictions:

$$\mathbb{E}(\mathbf{y}^*) \approx \frac{1}{T}\sum_{t=1}^{T} \widehat{\mathbf{y}}_t$$

- Use second moment for uncertainty (in regression):

$$\text{Var}(\mathbf{y}^*) \approx \frac{1}{T}\sum_{t=1}^{T} \widehat{\mathbf{y}}_t^T \widehat{\mathbf{y}}_t - \mathbb{E}(\mathbf{y}^*)^T \mathbb{E}(\mathbf{y}^*) + \tau^{-1}\mathbf{I}$$

with $\widehat{\mathbf{y}}_t \sim \text{DropoutNetwork}(\mathbf{x}^*)$.

In more practical terms, given point x:

- drop units at test time
- repeat 10 times
- and look at mean and sample variance.

Or in Python:

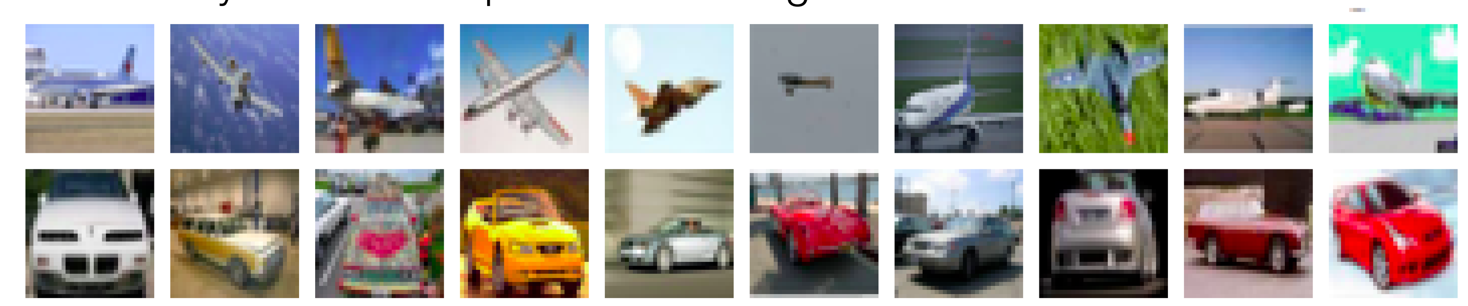
```
y = []
for _ in xrange(10):
    y.append(model.output(x, dropout=True))
y_mean = numpy.mean(y)
y_var = numpy.var(y)
```

Using the predictive mean

CIFAR-10 Test Error (and Std.)

Model	Standard Dropout	Bayesian technique
NIN	10.43 (Lin et al., 2013)	$\mathbf{10.27 \pm 0.05}$
DSN	9.37 (Lee et al., 2014)	$\mathbf{9.32 \pm 0.02}$
Augmented-DSN	7.95 (Lee et al., 2014)	$\mathbf{7.71 \pm 0.09}$

Bayesian techniques with existing CIFAR-10 state-of-the-art

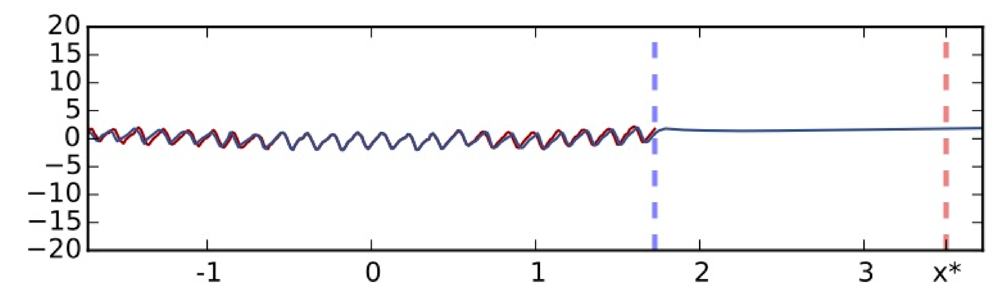


More results at `yarin.co/BCNN`

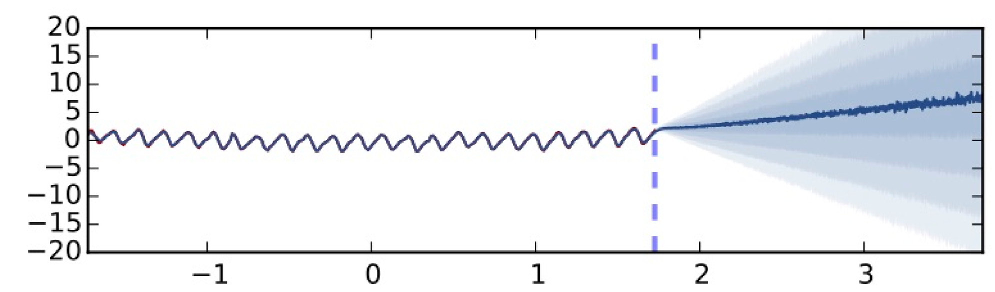
Using the predictive variance

What would be the CO_2 level in Mauna Loa, Hawaii, *in 20 years' time*?

- Normal dropout (weight averaging, 5 layers, ReLU units):



- Same network, Bayesian perspective:



Online demos at `yarin.co/blog`

And in numbers...

Dataset	Avg. Test RMSE and Std. Errors VI	PBP	Dropout	Avg. Test LL and Std. Errors VI	PBP	Dropout
Boston Housing	4.32 ±0.29	3.01 ±0.18	**2.97 ±0.85**	-2.90 ±0.07	-2.57 ±0.09	**-2.46 ±0.25**
Concrete Strength	7.19 ±0.12	5.67 ±0.09	**5.23 ±0.53**	-3.39 ±0.02	-3.16 ±0.02	**-3.04 ±0.09**
Energy Efficiency	2.65 ±0.08	1.80 ±0.05	**1.66 ±0.19**	-2.39 ±0.03	-2.04 ±0.02	**-1.99 ±0.09**
Kin8nm	**0.10 ±0.00**	**0.10 ±0.00**	**0.10 ±0.00**	0.90 ±0.01	0.90 ±0.01	**0.95 ±0.03**
Naval Propulsion	**0.01 ±0.00**	**0.01 ±0.00**	**0.01 ±0.00**	3.73 ±0.12	3.73 ±0.01	**3.80 ±0.05**
Power Plant	4.33 ±0.04	4.12 ±0.03	**4.02 ±0.18**	-2.89 ±0.01	-2.84 ±0.01	**-2.80 ±0.05**
Protein Structure	4.84 ±0.03	4.73 ±0.01	**4.36 ±0.04**	-2.99 ±0.01	-2.97 ±0.00	**-2.89 ±0.01**
Wine Quality Red	0.65 ±0.01	0.64 ±0.01	**0.62 ±0.04**	-0.98 ±0.01	-0.97 ±0.01	**-0.93 ±0.06**
Yacht Hydrodynamics	6.89 ±0.67	**1.02 ±0.05**	1.11 ±0.38	-3.43 ±0.16	-1.63 ±0.02	**-1.55 ±0.12**
Year Prediction MSD	9.034 ±NA	8.879 ±NA	**8.849 ±NA**	-3.622 ±NA	-3.603 ±NA	**-3.588 ±NA**

Table 1: **Average test performance in RMSE and predictive log likelihood** for a popular variational inference method (VI, Graves [20]), Probabilistic back-propagation (PBP, Hernández-Lobato and Adams [27]), and dropout uncertainty (Dropout).

More results at `yarin.co/dropout`

Full paper: "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning". Photos taken from Wikimedia or original work.