Calibrated and Sharp Uncertainties in Deep Learning via Density Estimation

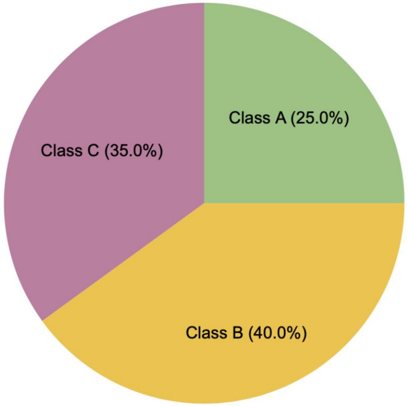
ICML 2022

Volodymyr Kuleshov, Shachi Deshpande

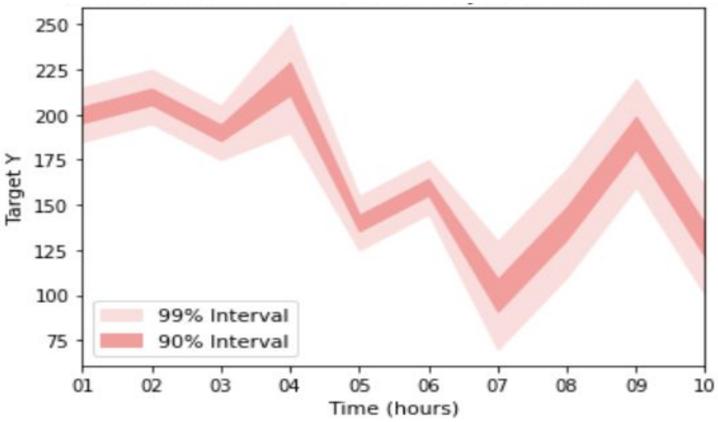
July 21, 2022

CORNELL TECH

Machine Learning Models Output Probabilistic Predictions



Class membership probabilities

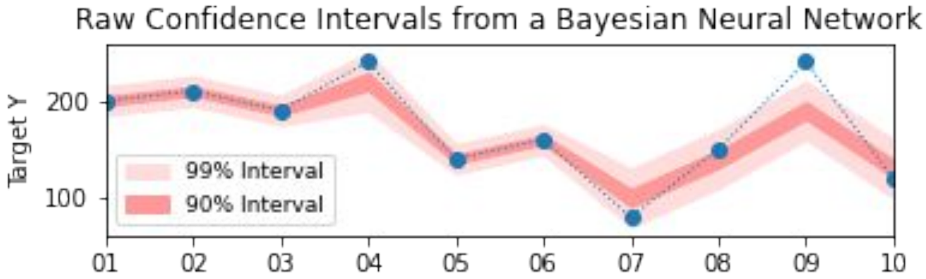


Raw confidence intervals from Bayesian Neural Network

CORNELL TECH

Machine Learning Models Output Probabilistic Predictions

Log-likelihood based training can give inaccurate probabilistic predictions!



Only **half** the points are within **90%** region!

Machine Learning Models Output Probabilistic Predictions

Accurate Probabilistic Forecast

Calibration

90% confidence interval contains the true outcome 90% times

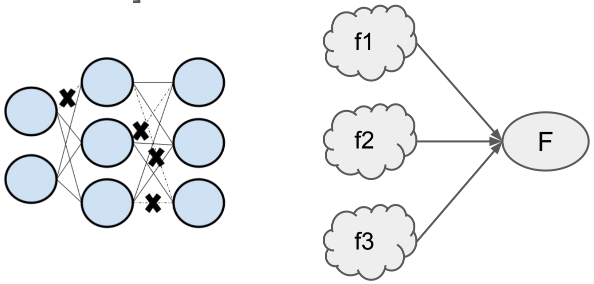
Sharpness

Narrow Confidence Intervals

Gneiting et al, 2007
Gneiting and Raftery, 2005; 2007

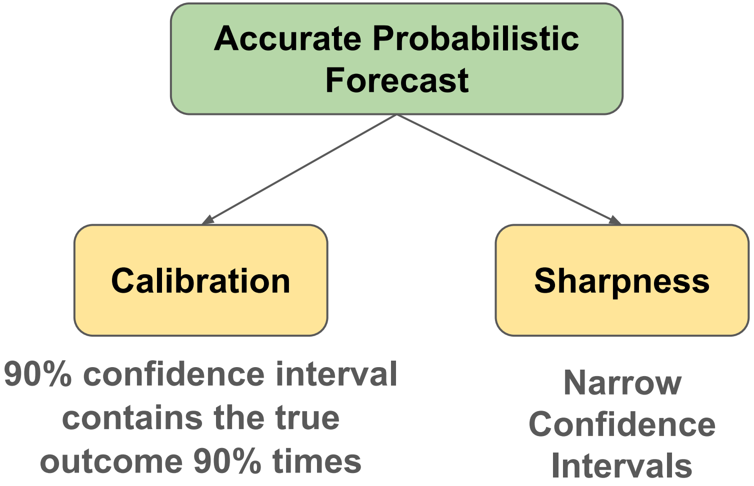
Solutions to improve probabilistic forecasts

- Dropout [Gal & Ghahramani, 2016] or ensembling [Lakshminarayanan et al,, 2017]



- Recalibration [Platt, 1999; Kuleshov et al. 2018; Song et al. 2019]

Calibration of Probabilistic Forecasts

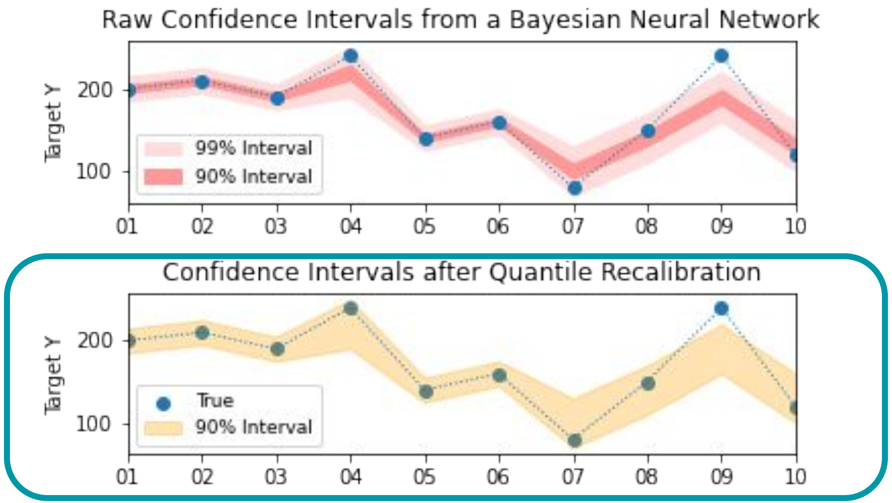


Gneiting et al, 2007
Gneiting and Raftery, 2005; 2007

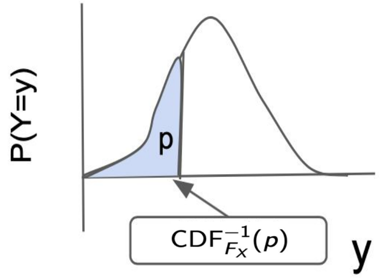
What we do

- **Simpler** techniques to obtain **distribution calibration** that are more broadly applicable
- Reason about uncertainty in terms of **calibration** and **sharpness**
- **Low-dimensional density estimation** to enforce distribution calibration
- **Theoretical analysis** to establish guarantees on calibration and vanishing regret

Quantile Calibration of Probabilistic Forecasts



Quantile Calibration



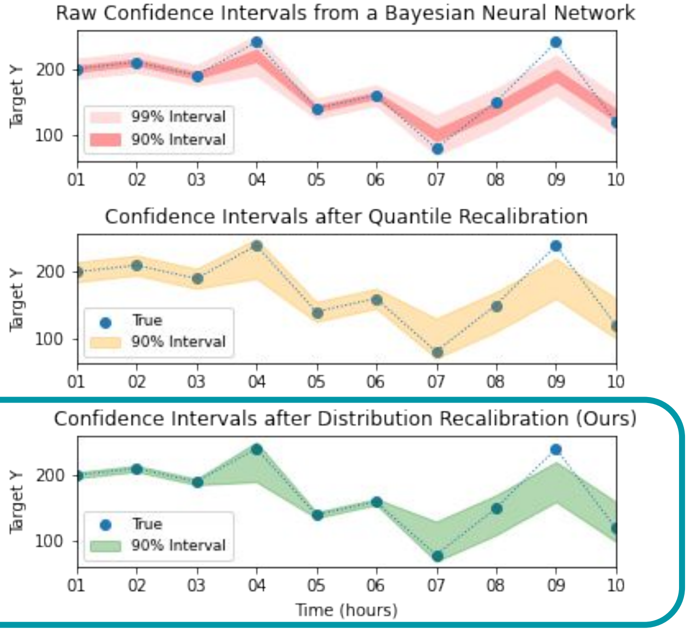
$$\mathbb{P}(Y \leq \mathrm{CDF}_{F_X}^{-1}(p)) = p \text{ for all } p \in [0,1],$$

9/10 points are within 90% region!

Quantile calibration re-labels 99% interval to 90% interval

CORNELL TECH

Distribution Calibration of Probabilistic Forecasts

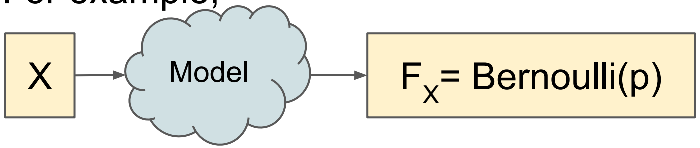


9/10 points are within 90% region and **tighter** confidence intervals after applying distribution calibration

Distribution Calibration

$\mathbb{P}(Y = y \mid F_X = F) = f(y)$ for all $y \in \mathcal{Y}, F \in \Delta_{\mathcal{Y}}$

For example,

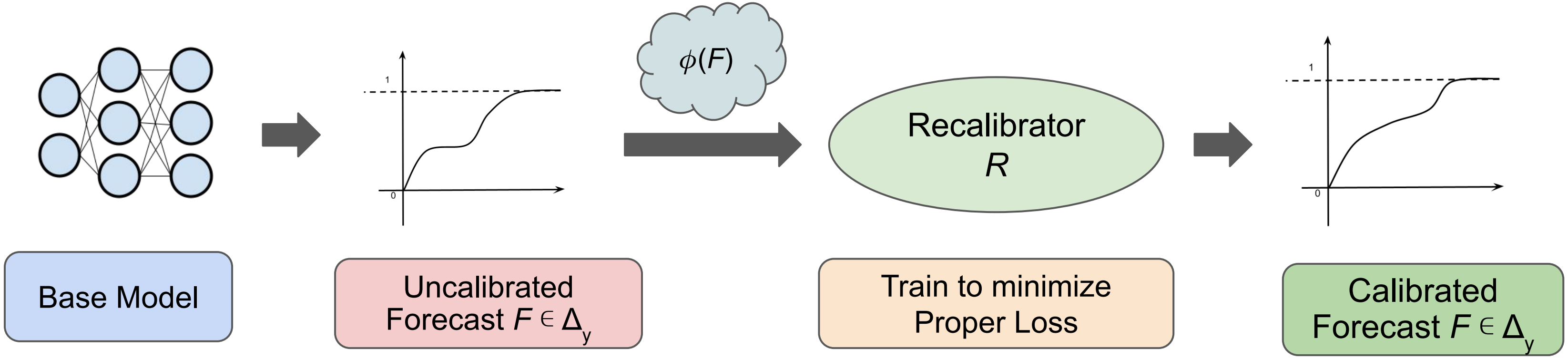


$P(Y=1|F_X=\text{Bern}(p)) = p$

Quantile Calibration

$\mathbb{P}(Y \leq \text{CDF}_{F_X}^{-1}(p)) = p$ for all $p \in [0, 1]$,

Training Distribution Calibrated Models



Recalibration as density estimation

$$R(F) = \mathbb{P}(Y \mid H(X) = F)$$

Challenge 1: Conditioning on *F*
(can be any arbitrary distribution)
Challenge 2: Learning objective for *R*

- Low dimensional density estimation (tractable)
- Proper Loss = **Calibration** - **Sharpness** + Irreducible term
- Quantile function regression when underlying model performs probabilistic regression
- Enforces stronger notion of calibration as compared to quantile calibration by Kuleshov et al. (2018)
- More broadly applicable as compared to distribution calibration by Song at al. (2019)

Theoretical Analysis and Experimental Results

- **Distribution Calibration**
  We prove that we can achieve distribution calibration via density estimation $P(Y| H(X)=F)$

- **Vanishing Regret**:
  We achieve calibration without degrading performance of baseline model (in terms of proper loss)

$$\text{ExpectedLoss}(R \circ H, Y) \leq \text{ExpectedLoss}(H, Y) + \delta_m$$

Experiments: UCI Regression benchmarks and Classification benchmarks
Key results:

- **Calibrated Regression**: Accuracies and uncertainties improve over Kuleshov et al. (2018) and in many cases over Song et al. (2019)
- **Calibrated Classification**: Best uncertainties are obtained via our method (compared with Platt scaling baseline) while accuracies remain similar

We test with a number of neural network base models. Please check our paper for more results!

Conclusions

- Reasoning about uncertainty in machine learning in terms of **calibration** and **sharpness**
- **Simple low-dimensional density estimation** to provably achieve distribution calibration
- Theoretical analysis to establish asymptotically distributionally calibrated forecasts while minimizing regret
- Calibration may be simpler than previously thought. Distribution calibration should be leveraged more broadly across machine learning

Thank you!