Lecture 2

Softmax as probabilities

- Scores transformed to probabilities

$$[1, -2, 0] \rightarrow [e^{1}, e^{-2}, e^{0}] = [2.71, 0.14, 1] \rightarrow [0.7, 0.04, 0.26]$$

- Scores scaled

$$[0.5, -1, 0] \rightarrow [e^{0.5}, e^{-1}, e^{0}] = [1.65, 0.37, 1] \rightarrow [0.55, 0.12, 0.33]$$

- Scaling?

Regularization

$$L = \underbrace{\frac{1}{N}\sum_i L_i}_{\text{data loss}} + \underbrace{\lambda R(W)}_{\text{regularization loss}}$$

- encode preference for a certain set of weights W over others
- As you vary \lambda different W will be chosen and will give different scores for an input
- Regularization important
  - Prevent overfitting
  - Selection of model
    - Bayesian interpretation

MAP vs MLE

- MLE
  - model is fixed (to be estimated) and data is random
  - "Prior" knowledge of model through regularization term

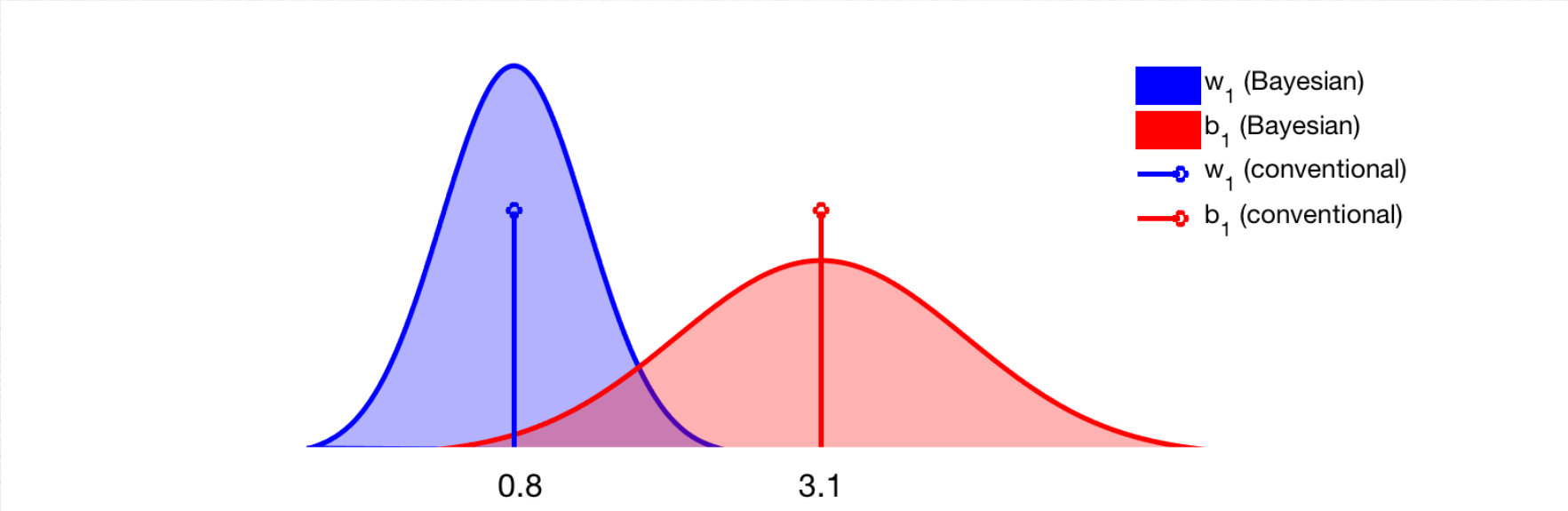
$$W_{ML} = argmax_W \Pi_{i=1}^{m} p(y_i \mid x_i; W)$$

$$W_{ML} = argmax_W (\Pi_{i=1}^{m} p(y_i \mid x_i; W) * \lambda e^{R(W)})$$

- MAP
  - Data is fixed, model is random and there is a prior probability on model

$$W_{MAP} = argmax_W \Pi_{i=1}^{m} p(y_i \mid x_i) p(W)$$

- Gaussian Prior p(W) will lead to a L2 regulariser



Having a distribution instead of a single value is a powerful thing. For one, it becomes possible to *sample* from a distribution *many many* times and see how this affect the predictions of the model. If it gives consistent predictions, sampling after sampling, then the net is said to be “confident” about its prediction.

Meaning of Confidence

- Bayesian Confidence Intervals
  - have a prior distribution over the way the world creates observations, use Bayes law to construct a posterior distribution over the way the world creates observations.
  - With respect to this posterior distribution, construct an interval containing the truth with high probability.
  - The semantics of a Bayesian confidence interval is “If the world is drawn from the prior the interval contains the truth with high probability”.
  - No assumption of independent samples is required.
  - Unlike classical confidence intervals, it’s easy to have a statement conditioned on features.
    - For example, “the probability of disease given the observations is in [0.8,1]”.

Contd

- Confidence== Probability
  - If probability = 0.5 is it because learning algorithm has not seen this event, or is this event uncertain?
  - Also we need to make sure we have probabilities from learning algorithm scores (Calibration)
- Confidence Intervals
  - True but hidden value — construct an interval around this hidden value

Ensembles

- Simplest Ensemble that also has a Bayesian formulation
  - MC-Dropout Network
    - Dropout a mechanism for regularising NN while training
      - p neurons are turned off during training; so every sample (or batch) is seeing a different network
      - Conventionally all neurons are used while inferencing (with appropriate scaling)
    - MC-Dropout network turns off p neurons while inferencing also. Uses multiple (T) such networks to make the inference
- Other possibilities for this ensemble methods also (NIPS 2017)

T ensembles

- Given x input, output is $y^*=E(y/x)$ for each model in ensemble
- Predicted output is mean over ensemble and confidence is variance over ensembles.

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

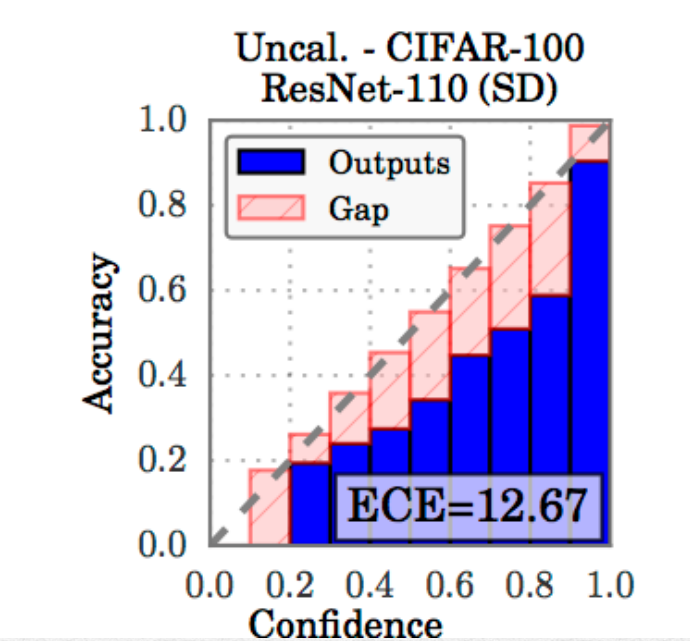
$$Var(y^*) \approx \tau^{-1} I_D + \frac{1}{T}\sum_{t=1}^{T} \hat{y}_t^*(x^*)^T \hat{y}_t^*(x^*) - \mathbb{E}(y^*)^T \mathbb{E}(y^*)$$

Calibration

- frequentist notion of uncertainty which measures the discrepancy between subjective forecasts and(empirical) long-run frequencies.
- The quality of calibration can be measured by proper scoring rules
- Note that calibration is an orthogonal concern to accuracy: a network's predictions may be accurate and yet miscalibrated, and vice versa.

Measure Miscalibration

- Given a set of predictions (on validation data) with associated probabilities (confidence)
  - Bin the confidence in N bins
  - For each bin the average confidence should be close to average accuracy for that bin
  - Difference between accuracy of bin and confidence of bin is referred to as calibration gap
  - A weighted (with number of samples in bin) average of above number is the Expected Calibration Error (ECE)



Calibrate

- Postprocessing
- Platt scaling
  - Initially applied to SVM scores but can be applied to logistic regression also
  - *f(x)* is the regression function we are learning.
  - The non-probabilistic predictions of a classifier are used as features for a logistic regression model, which is trained on the validation set to return probabilities
  - The parameters A and B are estimated using a maximum likelihood method that optimizes on the same training set as that for the original classifier *f*.
  - To avoid overfitting to this set, a held-out calibration set or cross-validation can be used
- Temperature Scaling
  - A new form where the above transformation is simpler (only one parameter)
    - T large — move towards maximum uncertainty
    - T = 1 — original uncertainty
    - T = 0 probability collapses to point mass, no uncertainty

$$P(y = 1 \mid x) = \frac{1}{1 + exp(Af(x) + B}$$

$$P(y \mid x) = \frac{e^{(z/T)}}{\sum_i e^{(z_i/T)}}$$