Bayesian inference is suboptimal for learning under model misspecification

Andrés R. Masegosa
University of Almería

Abstract

- Most of machine learning models are misspecified.
- A novel PAC-Bayesian analysis shows that Bayesian model averaging is suboptimal for generalization under misspecification.
- A novel learning framework explicitly addressing misspecification is presented.

The learning problem

- **Assumptions**:
  - $\nu(x)$ is the data generating distribution (**unknown**).
  - **Model Misspecification**: $\forall\theta\ p(\cdot|\theta)\neq\nu$.
- The ***predictive posterior distribution*** for a given $\rho(\theta)$,

$$p(x)=\int p(x|\theta)\rho(\theta)d\theta=\mathbb{E}_\rho[p(x|\theta)]$$

- **The learning problem** is defined as,

$$\rho^\star=\arg\min_\rho KL(\nu(x),\mathbb{E}_\rho[p(x|\theta)])=\arg\min_\rho \underbrace{\mathbb{E}_{\nu(x)}[-\ln\mathbb{E}_\rho[p(x|\theta)]]}_{CE(\rho)}$$

- $CE(\rho)$ measures the **generalization error** associated to ρ.

First-order PAC-Bayes bounds and the Bayesian posterior

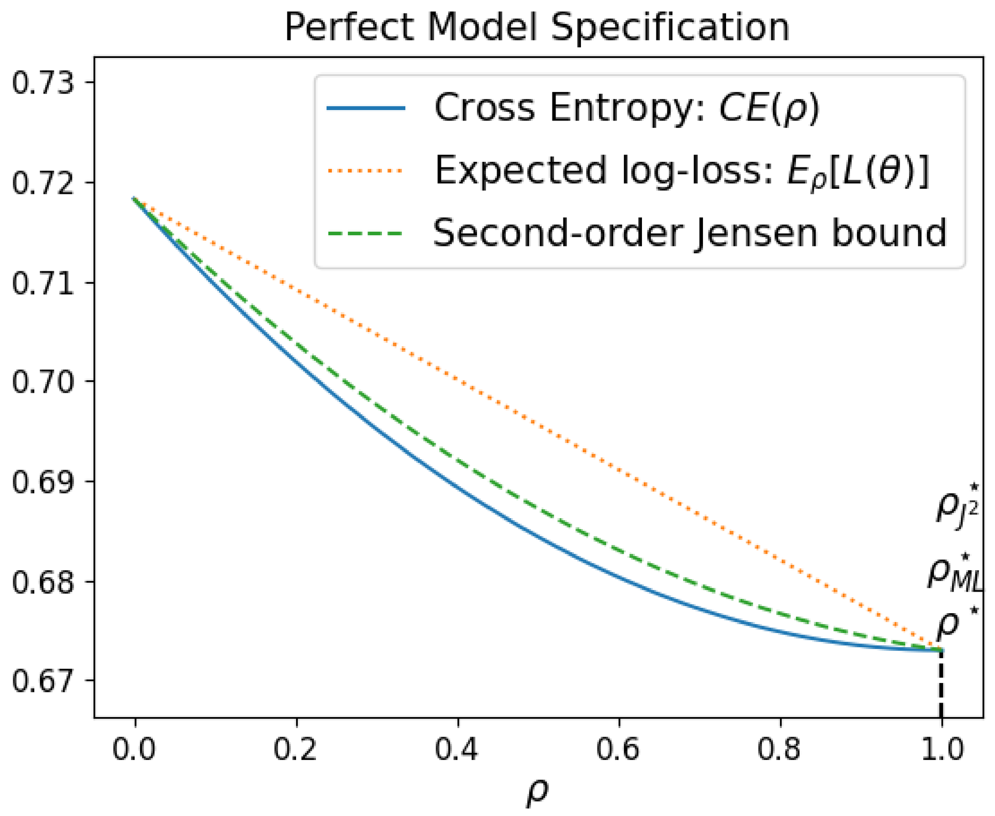
Germain et al. 2016 showed the Bayesian posterior minimize a (first-order) PAC-Bayesian bound:

$$\underbrace{p(\theta|D)}_{\text{Bayesian Posterior}}=\arg\min_\rho \underbrace{\mathbb{E}_\rho[\overbrace{\hat{L}(\theta,D)}^{\text{Empirical log-loss}}]+\frac{KL(\rho,\pi)}{n}+cte}_{\text{First-Order PAC-Bayes bound}}$$

First-order PAC-Bayes upper bounds the generalization error $CE(\rho)$

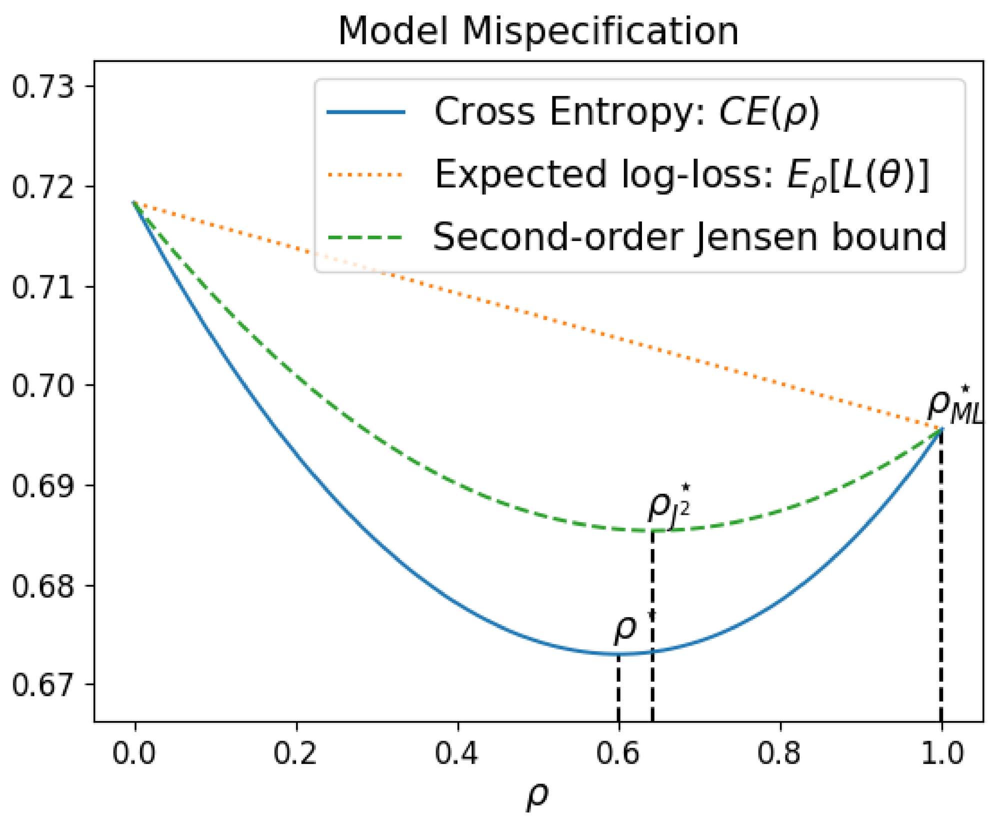
$$\underbrace{CE(\rho)}_{\text{Generalization Error}}\leq\underbrace{\mathbb{E}_\rho[L(\theta)]}_{\text{Jensen bound}}\leq\underbrace{\mathbb{E}_\rho[\hat{L}(\theta,D)]+\frac{KL(\rho,\pi)}{n}+cte}_{\text{First-Order PAC-Bayes bound}}$$

Bayesian posterior is optimal under perfect specification



$$\underbrace{CE(\rho)}_{\text{Generalization Error}}\leq\underbrace{\mathbb{E}_\rho[L(\theta)]}_{\text{Jensen bound}}\leq\underbrace{\mathbb{E}_\rho[\hat{L}(\theta,D)]+\frac{KL(\rho,\pi)}{n}+cte}_{\text{First-Order PAC-Bayes bound}}$$

Bayesian posterior is not optimal under model misspecification



$$\underbrace{CE(\rho)}_{\text{Generalization Error}}\leq\underbrace{\mathbb{E}_\rho[L(\theta)]-\overbrace{\mathbb{V}(\rho)}^{\text{Variance}}}_{\text{Second-order Jensen bound}}\leq\underbrace{\mathbb{E}_\rho[\hat{L}(\theta,D)]-\hat{\mathbb{V}}(\rho,D)+\frac{KL(\rho,\pi)}{n}+cte}_{\text{Second-order PAC-Bayes bound}}$$

A new learning framework

Minimizing second-order PAC-Bayes bounds

$$\arg\min_{\rho\in Q}\mathbb{E}_\rho[L(\theta,D)]-\hat{\mathbb{V}}(\rho,D)+\frac{KL(\rho,\pi)}{n}+cte$$

where Q is a tractable family of densities.

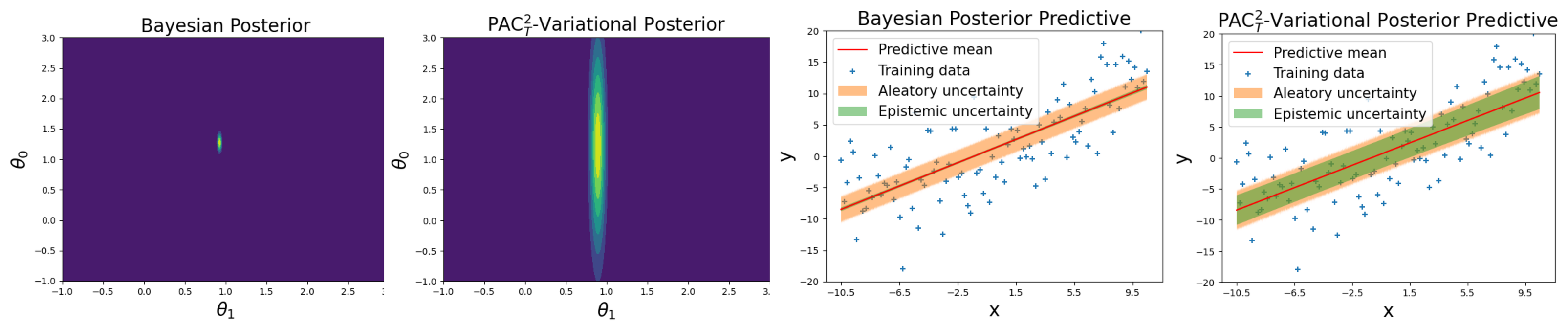


Figure 1: Bayesian Linear Regression.

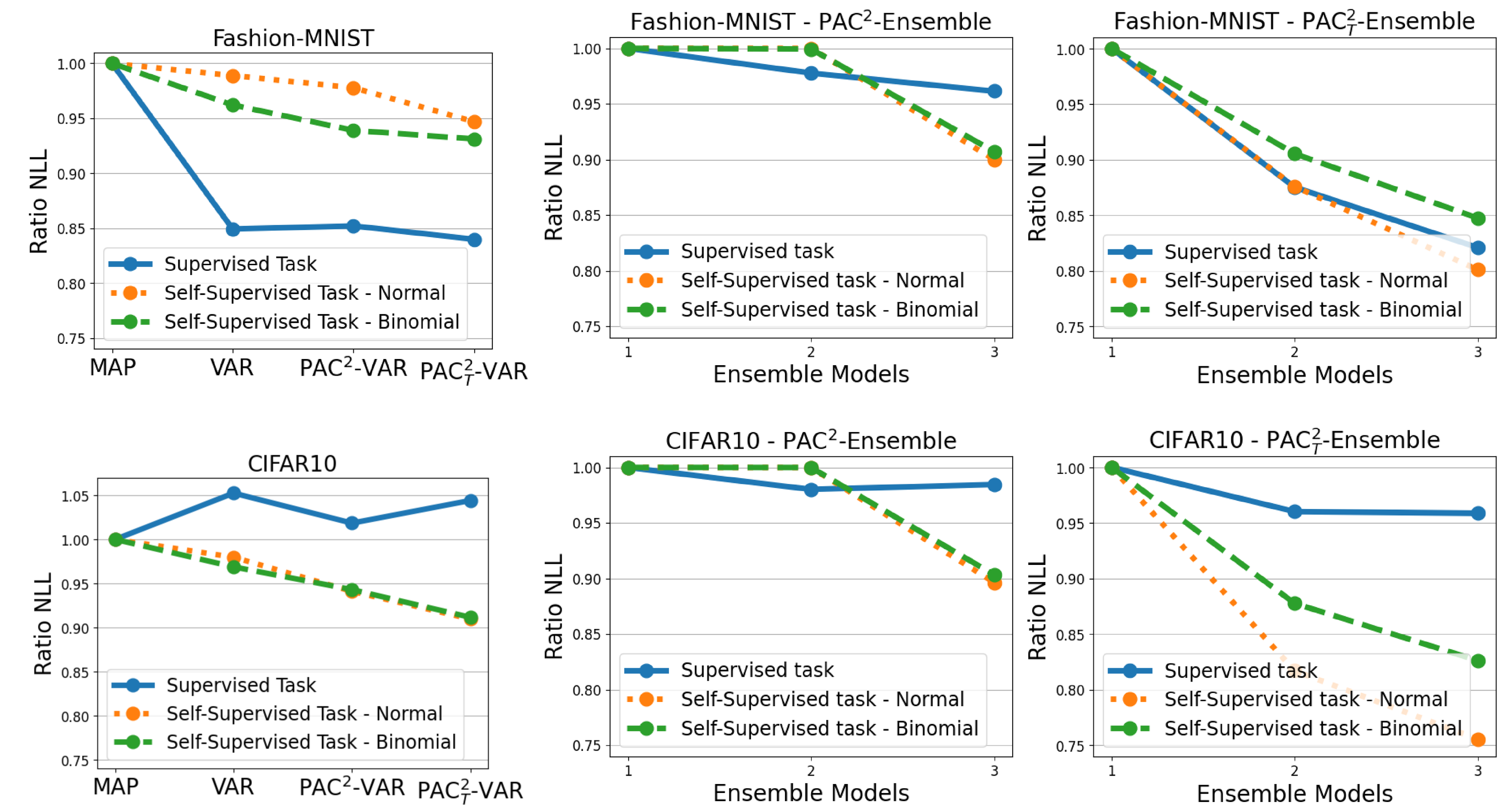


Figure 2: Bayesian Neural Networks.

Summary

- Bayesian methods are suboptimal for learning predictive models when the model family is misspecified.
- Second order PAC-Bayes bounds directly address model misspecification when learning.