Bayesian model averaging is suboptimal for generalization under model misspecification

Andrés Masegosa

Department of Mathematics
University of Almería
Spain

The learning problem

- **Notation**:
  - $\nu(\mathbf{x})$ is the data generating distribution (**unknown**).
  - $p(\mathbf{x}|\boldsymbol{\theta})$ is a probabilistic model parametrized by $\boldsymbol{\theta}$.
  - $\forall\boldsymbol{\theta}\ \nu(\mathbf{x}) \neq p(\mathbf{x}|\boldsymbol{\theta})$.

The learning problem

- **Notation**:
  - $\nu(\mathbf{x})$ is the data generating distribution (**unknown**).
  - $p(\mathbf{x}|\boldsymbol{\theta})$ is a probabilistic model parametrized by $\boldsymbol{\theta}$.
  - $\forall\boldsymbol{\theta}\ \nu(\mathbf{x}) \neq p(\mathbf{x}|\boldsymbol{\theta})$.

- The **predictive posterior distribution** for a given $\rho(\boldsymbol{\theta})$,

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

- **Notation**:
  - $\nu(\mathbf{x})$ is the data generating distribution (**unknown**).
  - $p(\mathbf{x}|\boldsymbol{\theta})$ is a probabilistic model parametrized by $\boldsymbol{\theta}$.
  - $\forall\boldsymbol{\theta}\ \nu(\mathbf{x}) \neq p(\mathbf{x}|\boldsymbol{\theta})$.
- The **predictive posterior distribution** for a given $\rho(\boldsymbol{\theta})$,

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

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

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

- **Notation**:
  - $\nu(\mathbf{x})$ is the data generating distribution (**unknown**).
  - $p(\mathbf{x}|\boldsymbol{\theta})$ is a probabilistic model parametrized by $\boldsymbol{\theta}$.
  - $\forall\boldsymbol{\theta}\ \nu(\mathbf{x}) \neq p(\mathbf{x}|\boldsymbol{\theta})$.

- The **predictive posterior distribution** for a given $\rho(\boldsymbol{\theta})$,

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

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

$$\rho^{\star} = \arg\min_{\rho} KL(\nu(\mathbf{x}), \mathbb{E}_{\rho(\theta)}[p(\mathbf{x}|\boldsymbol{\theta})]) = \arg\min_{\rho} \underbrace{\mathbb{E}_{\nu(\mathbf{x})}[\ln \frac{1}{\mathbb{E}_{\rho(\theta)}[p(\mathbf{x}|\boldsymbol{\theta})]}]}_{CE(\rho)}$$

- **Notation**:
  - $\nu(\mathbf{x})$ is the data generating distribution (**unknown**).
  - $p(\mathbf{x}|\boldsymbol{\theta})$ is a probabilistic model parametrized by $\boldsymbol{\theta}$.
  - $\forall\boldsymbol{\theta}\ \nu(\mathbf{x}) \neq p(\mathbf{x}|\boldsymbol{\theta})$.

- The **predictive posterior distribution** for a given $\rho(\boldsymbol{\theta})$,

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

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

$$\rho^{\star} = \arg\min_{\rho} KL(\nu(\mathbf{x}), \mathbb{E}_{\rho(\theta)}[p(\mathbf{x}|\boldsymbol{\theta})]) = \arg\min_{\rho} \underbrace{\mathbb{E}_{\nu(\mathbf{x})}[\ln \frac{1}{\mathbb{E}_{\rho(\theta)}[p(\mathbf{x}|\boldsymbol{\theta})]}]}_{CE(\rho)}$$

Learning from a finite dataset

- We do not have access to $\nu(\mathbf{x})$, only to a i.i.d. sample $D = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$.

First-order PAC-Bayes bounds

Remind!

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

Remind!

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

$$CE(\rho) \overset{Jensen}{\leq} \underbrace{\mathbb{E}_{\rho(\theta)}[L(\boldsymbol{\theta})]}_{Jensen\ bound}$$

Remind!

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

$$CE(\rho) \overset{Jensen}{\leq} \underbrace{\mathbb{E}_{\rho(\theta)}[L(\boldsymbol{\theta})]}_{Jensen\ bound} \overset{PAC-Bayes}{\lesssim} \underbrace{\mathbb{E}_{\rho(\theta)}[L(\boldsymbol{\theta}, D)] + \frac{KL(\rho, \pi) + \ln\frac{1}{\xi} + \psi_{\pi,\nu}(c, n)}{cn}}_{PAC-Bayes\ bound}$$

Remind!

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

$$CE(\rho) \overset{Jensen}{\leq} \underbrace{\mathbb{E}_{\rho(\theta)}[L(\boldsymbol{\theta})]}_{Jensen\ bound} \overset{PAC-Bayes}{\lesssim} \underbrace{\mathbb{E}_{\rho(\theta)}[L(\boldsymbol{\theta}, D)] + \frac{KL(\rho,\pi) + \ln\frac{1}{\xi} + \psi_{\pi,\nu}(c,n)}{cn}}_{PAC-Bayes\ bound}$$

The learning strategy is to minimize the PAC-Bayes bound

- $\rho^\star$ is the **Bayesian posterior** for $c = 1$ (Germain et al. 2016),

$$\rho^\star = p(\boldsymbol{\theta}|D) = \frac{p(D|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int p(D|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}$$

- The **Bayesian learning strategy**,

$$CE(\rho) \overset{Jensen}{\leq} \underbrace{\mathbb{E}_{\rho(\theta)}[L(\boldsymbol{\theta})]}_{Jensen\ bound} \overset{PAC-Bayes}{\lesssim} \underbrace{\mathbb{E}_{\rho(\theta)}[L(\boldsymbol{\theta}, D)] + \frac{KL(\rho, \pi) + \ln\frac{1}{\xi} + \psi_{\pi,\nu}(c, n)}{cn}}_{PAC-Bayes\ bound}$$

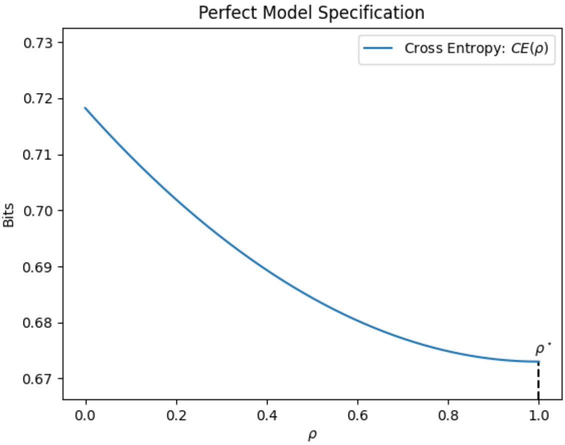
- The **Bayesian learning strategy**,

$$CE(\rho) \overset{Jensen}{\leq} \underbrace{\mathbb{E}_{\rho(\theta)}[L(\boldsymbol{\theta})]}_{Jensen\ bound} \overset{PAC-Bayes}{\lesssim} \underbrace{\mathbb{E}_{\rho(\theta)}[L(\boldsymbol{\theta}, D)] + \frac{KL(\rho, \pi) + \ln\frac{1}{\xi} + \psi_{\pi,\nu}(c, n)}{cn}}_{PAC-Bayes\ bound}$$

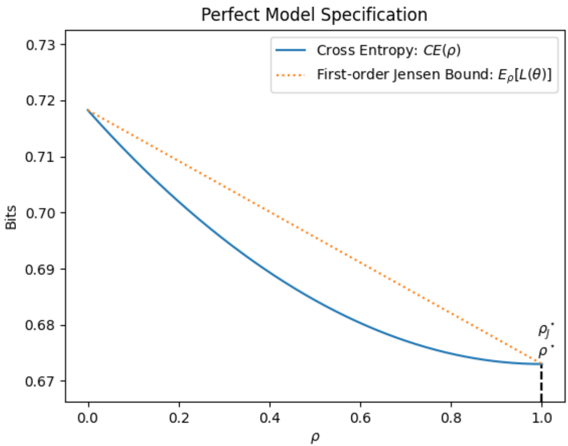
- The **minimum of the Jensen bound is a Dirac-delta distribution** centered around,

$$\boldsymbol{\theta}^{\star}_{ML} = \arg\min_{\theta} KL(\nu(\mathbf{x}), p(\mathbf{x}|\boldsymbol{\theta}))$$

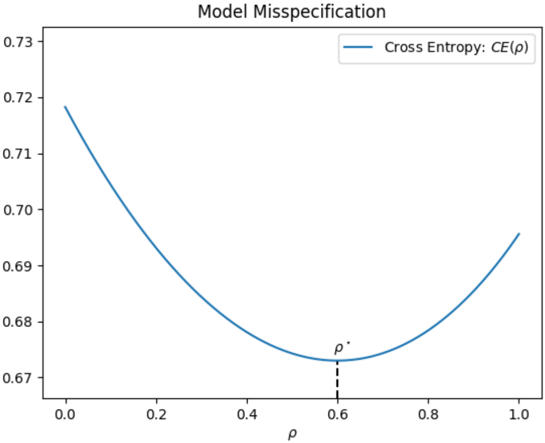
Is the Bayesian approach an optimal learning strategy?



$$CE(\rho)$$

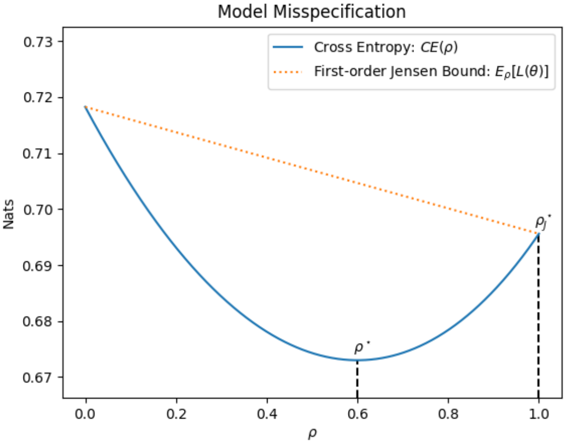


$$CE(\rho) \overset{Jensen}{\leq} \underbrace{\mathbb{E}_{\rho(\theta)}[L(\boldsymbol{\theta})]}_{Jensen\ bound}$$


Model Misspecification
Cross Entropy: $CE(\rho)$
0.73
0.72
0.71
0.70
0.69
0.68
0.67
ρ^*
0.0
0.2
0.4
0.6
0.8
1.0
ρ

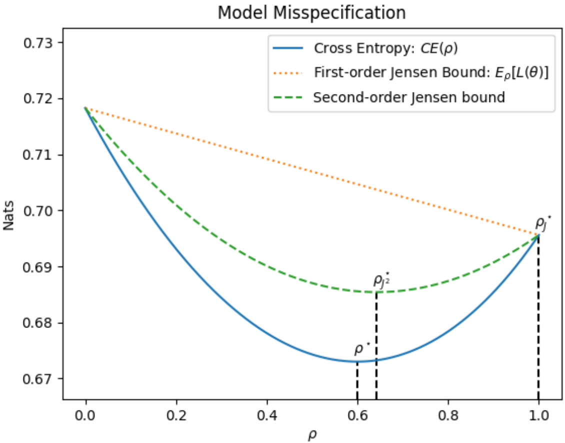
$$CE(\rho)$$

Is the Bayesian approach an optimal learning strategy?



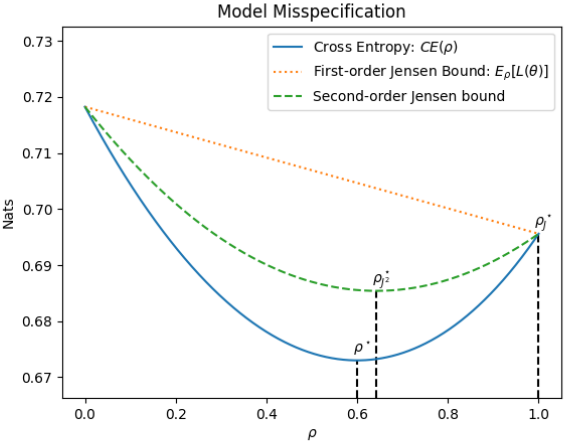
$$CE(\rho) \overset{Jensen}{\leq} \underbrace{\mathbb{E}_{\rho(\theta)}[L(\boldsymbol{\theta})]}_{Jensen\ bound}$$

Is the Bayesian approach an optimal learning strategy?



$$CE(\rho) \overset{(Liao\&Berg, 2019)}{\leq} \underbrace{\mathbb{E}_{\rho(\theta)}[L(\boldsymbol{\theta})] - \mathbb{V}(\rho)}_{Second-order\ Jensen\ bound}$$

Is the Bayesian approach an optimal learning strategy?



$$\underbrace{\mathbb{E}_{\rho(\theta)}[L(\boldsymbol{\theta})] - \mathbb{V}(\rho)}_{Second-order\ Jensen\ bound} \overset{PAC-Bayes}{\lesssim} \underbrace{\mathbb{E}_{\rho(\theta)}[L(\boldsymbol{\theta}, D)] - \hat{\mathbb{V}}(\rho, D) + \frac{KL(\rho,\pi) + \ln\frac{1}{\xi} + \psi_{\pi,\nu}(c,n)}{cn}}_{Second-order\ PAC-Bayes\ bound}$$

Minimizing second-order PAC-Bayes bound

- Variational methods for minimizing **second-order PAC-Bayes bounds**,

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

where Q is a tractable family of densities (i.e. fully factorized Gaussian distribution).

A new learning framework

Minimizing second-order PAC-Bayes bound

- Variational methods for minimizing **second-order PAC-Bayes bounds**,

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

where Q is a tractable family of densities (i.e. fully factorized Gaussian distribution).

Variational Inference

- **Standard Variational methods** tries to minimize the first-order PAC-Bayes bound,

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

Conclusions

- The Bayesian approach does **not** seem to be **an optimal learning strategy**.
- Novel **variational and ensemble learning algorithms**.

https://github.com/PGM-Lab/PAC2BAYES