# Learning under model misspecification

## Applications to Variational and Ensemble methods

Andrés R. Masegosa

*University of Almería*
*Spain*

November 25, 2020

# Introduction

## Model Misspecification

- Virtually any model we use **does not perfectly represent reality**.
- We mostly work in the **model misspecification** regime.

# Abstract

## Model Misspecification

- Virtually any model we use **does not perfectly represent reality**.
- We mostly work in the **model misspecification** regime.

## Contributions

- **Generalization** analysis of Bayesian learning under model misspecification.
- Bayesian model averaging is **suboptimal** for generalization.
- **New learning framework** which explicitly addresses model misspecfication.
- Empirical evaluations on **Bayesian deep learning** illustrate this approach.

## Assumption 1: I.I.D. Data

- There exists an **underlying distribution** $\nu(\mathbf{x})$ generating the training/test data.
- The **training data sample**, $D = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, is i.i.d. from $\nu(\mathbf{x})$.

## Assumption 2: Model misspecification

- Our model class only **approximates** reality (not prefect).
- $p(\mathbf{x}|\boldsymbol{\theta})$ is our (parametric) **probabilistic model class**.

## Assumption 2: Model misspecification

- Our model class only **approximates** reality (not prefect).
- $p(\mathbf{x}|\boldsymbol{\theta})$ is our (parametric) **probabilistic model class**.

$$\forall \boldsymbol{\theta} \in \boldsymbol{\Theta} \quad \nu \neq p(\cdot|\boldsymbol{\theta})$$

## Assumption 3: Likelihood is Upper-Bounded

- There exists a $M > 0$

$$\forall \mathbf{x} \in \mathcal{X},\ \forall \boldsymbol{\theta} \in \boldsymbol{\Theta} \quad p(\cdot|\boldsymbol{\theta}) \leq M,$$

## Assumption 3: Likelihood is Upper-Bounded

- There exists a $M > 0$

$$\forall \mathbf{x} \in \mathcal{X},\ \forall \boldsymbol{\theta} \in \boldsymbol{\Theta} \quad p(\cdot|\boldsymbol{\theta}) \leq M,$$

- It holds in **supervised classification** (e.g. $M = 1$) and it may require to constrain the parameter space (e.g. the variance of the Gaussian higher than $\epsilon > 0$),.

# The Learning Problem

# The learning problem

- **Notation**: $\rho(\boldsymbol{\theta})$ is a probability distribution over the parameters of my model.

# The learning problem

- **Notation**: $\rho(\boldsymbol{\theta})$ is a probability distribution over the parameters of my model.

- 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**: $\rho(\boldsymbol{\theta})$ is a probability distribution over the parameters of my model.
- 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,

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

- **Notation**: $\rho(\boldsymbol{\theta})$ is a probability distribution over the parameters of my model.
- 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,

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

- ... is equivalent to:

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

- **Notation**: $\rho(\boldsymbol{\theta})$ is a probability distribution over the parameters of my model.
- 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,

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

- ... is equivalent to:

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

- $CE(\rho)$ measures the **generalization error** (or the predicitive risk) associated to $\rho$.

# The learning strategy

- How to solve this problem

$$\rho^{\star} = \arg\min_{\rho} \underbrace{CE(\rho)}_{\substack{\text{Generalization} \\ \text{Error}}}$$

if we do not have access to $\nu(\mathbf{x})$

- How to solve this problem

$$\rho^{\star} = \arg\min_{\rho} \underbrace{CE(\rho)}_{\text{Generalization Error}}$$

if we do not have access to $\nu(\mathbf{x})$

## The learning strategy

- The solution is to **employ upper-bounds**:

$$CE(\rho) \underbrace{\leq}_{\text{Jensen inequality}} \text{Oracle-Bound}(\rho) \underbrace{\lesssim}_{\text{w.p. } (1-\xi)} \text{Empirical-Bound}(\rho, D, \xi)$$

- How to solve this problem

$$\rho^\star = \arg\min_\rho \underbrace{CE(\rho)}_{\text{Generalization Error}}$$

if we do not have access to $\nu(\mathbf{x})$

## The learning strategy

- The solution is to **employ upper-bounds**:

$$CE(\rho) \underbrace{\leq}_{\text{Jensen inequality}} \text{Oracle-Bound}(\rho) \underbrace{\lesssim}_{\text{w.p. } (1-\xi)} \text{Empirical-Bound}(\rho, D, \xi)$$

- ... and **minimize** Empirical-Bound$(\rho, D, \xi)$,

$$\min_\rho \text{Empirical-Bound}(\rho, D, \xi)$$

- How to solve this problem

$$\rho^\star = \arg\min_\rho \underbrace{CE(\rho)}_{\text{Generalization Error}}$$

if we do not have access to $\nu(\mathbf{x})$

## The learning strategy

- The solution is to **employ upper-bounds**:

$$CE(\rho) \underbrace{\leq}_{\text{Jensen inequality}} \text{Oracle-Bound}(\rho) \underbrace{\lesssim}_{\text{w.p. } (1-\xi)} \text{Empirical-Bound}(\rho, D, \xi)$$

- ... and **minimize** Empirical-Bound$(\rho, D, \xi)$,

$$\min_\rho \text{Empirical-Bound}(\rho, D, \xi)$$

- The quality of the solution is going to depend of the **quality of the bounds**.

# First-order Jensen bounds and the Bayesian posterior

$$\underbrace{CE(\rho)}_{\text{Generalization Error}} \overbrace{\leq}^{\text{Jensen Inequality}} \underbrace{\mathbb{E}_\rho[L(\boldsymbol{\theta})]}_{\text{Oracle bound}}$$

$L(\boldsymbol{\theta})$ is the **expected log-loss**, $L(\boldsymbol{\theta}) = -\mathbb{E}_{\nu(\mathbf{x})}[\ln p(\mathbf{x}|\boldsymbol{\theta})]$.

$$\underbrace{CE(\rho)}_{\text{Generalization Error}} \overbrace{\leq}^{\text{Jensen Inequality}} \underbrace{\mathbb{E}_\rho[L(\boldsymbol{\theta})]}_{\text{Oracle bound}} \overbrace{\lesssim}^{\text{w.p. } (1-\xi)} \underbrace{\mathbb{E}_\rho[\hat{L}(\boldsymbol{\theta}, D)] + \frac{KL(\rho, \pi)}{n} + cte}_{\text{PAC-Bayes bound (Alquier et al. 2016)}}$$

$L(\boldsymbol{\theta})$ is the **expected log-loss**, $L(\boldsymbol{\theta}) = -\mathbb{E}_{\nu(\mathbf{x})}[\ln p(\mathbf{x}|\boldsymbol{\theta})]$.

$\hat{L}(\boldsymbol{\theta}, D)$ is the **empirical log-loss**, $L(\boldsymbol{\theta}, D) = -\frac{1}{n} \ln p(D|\boldsymbol{\theta})$.

## The Bayesian posterior (Germain et al. 2016)

- The learning strategy is to **minimize the PAC-Bayes bound**,

$$\rho^\star \quad = \quad \arg\min_{\rho} \underbrace{\mathbb{E}_{\rho(\theta)}[L(\boldsymbol{\theta}, D)] + \frac{KL(\rho, \pi)}{n} + cte}_{\text{PAC-Bayes bound (Alquier et al. 2016)}}$$

## The Bayesian posterior (Germain et al. 2016)

- The learning strategy is to **minimize the PAC-Bayes bound**,

$$\rho^\star = \arg\min_\rho \underbrace{\mathbb{E}_{\rho(\theta)}[L(\boldsymbol{\theta}, D)] + \frac{KL(\rho, \pi)}{n} + cte}_{\text{PAC-Bayes bound (Alquier et al. 2016)}}$$

$$= \arg\max_\rho \underbrace{\mathbb{E}_{\rho(\theta)}[\ln p(D|\boldsymbol{\theta})] - KL(\rho, \pi)}_{\text{Evidence Lower Bound (ELBO)}}$$

## The Bayesian posterior (Germain et al. 2016)

- The learning strategy is to **minimize the PAC-Bayes bound**,

$$\rho^\star = \arg\min_\rho \underbrace{\mathbb{E}_{\rho(\theta)}[L(\boldsymbol{\theta}, D)] + \frac{KL(\rho, \pi)}{n} + cte}_{\text{PAC-Bayes bound (Alquier et al. 2016)}}$$

$$= \arg\max_\rho \underbrace{\mathbb{E}_{\rho(\theta)}[\ln p(D|\boldsymbol{\theta})] - KL(\rho, \pi)}_{\text{Evidence Lower Bound (ELBO)}}$$

- $\rho^\star$ is the **Bayesian posterior**,

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

## The Bayesian posterior (Germain et al. 2016)

- The learning strategy is to **minimize the PAC-Bayes bound**,

$$\rho^\star = \arg\min_\rho \underbrace{\mathbb{E}_{\rho(\theta)}[L(\boldsymbol{\theta}, D)] + \frac{KL(\rho, \pi)}{n} + cte}_{\text{PAC-Bayes bound (Alquier et al. 2016)}}$$

$$= \arg\max_\rho \underbrace{\mathbb{E}_{\rho(\theta)}[\ln p(D|\boldsymbol{\theta})] - KL(\rho, \pi)}_{\text{Evidence Lower Bound (ELBO)}}$$

- $\rho^\star$ is the **Bayesian posterior**,

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

## The Bayesian posterior is a proxy

$$p(\boldsymbol{\theta}|D) \approx \arg\min_\rho KL\Big( \underbrace{\nu(\mathbf{x})}_{\substack{\text{Data} \\ \text{distribution}}} , \underbrace{\mathbb{E}_{\rho(\theta)}[p(\mathbf{x}|\boldsymbol{\theta})]}_{\substack{\text{Predictive} \\ \text{posterior}}} \Big)$$

# Is the Bayesian approach an optimal learning strategy?

- The **Bayesian learning strategy**,

$$\underbrace{CE(\rho)}_{\text{Generalization Error}} \overbrace{\leq}^{\text{Jensen Inequality}} \underbrace{\mathbb{E}_\rho[L(\boldsymbol{\theta})]}_{\text{First-Order Jensen bound}} \overbrace{\lesssim}^{\text{w.p. } (1-\xi)} \underbrace{\mathbb{E}_\rho[\hat{L}(\boldsymbol{\theta}, D)] + \frac{KL(\rho, \pi)}{n} + cte}_{\text{PAC-Bayes bound (Alquier et al. 2016)}}$$

# Is the Bayesian approach an optimal learning strategy?

- The **Bayesian learning strategy**,

$$\underbrace{CE(\rho)}_{\text{Generalization Error}} \overbrace{\leq}^{\text{Jensen Inequality}} \underbrace{\mathbb{E}_\rho[L(\boldsymbol{\theta})]}_{\text{First-Order Jensen bound}} \overbrace{\lesssim}^{\text{w.p. } (1-\xi)} \underbrace{\mathbb{E}_\rho[\hat{L}(\boldsymbol{\theta}, D)] + \frac{KL(\rho, \pi)}{n} + cte}_{\text{PAC-Bayes bound (Alquier et al. 2016)}}$$

- The Bayesian posterior **converges** to the minimum of $\mathbb{E}_\rho[L(\boldsymbol{\theta})]$.

# Is the Bayesian approach an optimal learning strategy?

- The **Bayesian learning strategy**,

$$\underbrace{CE(\rho)}_{\text{Generalization Error}} \overbrace{\leq}^{\text{Jensen Inequality}} \underbrace{\mathbb{E}_\rho[L(\boldsymbol{\theta})]}_{\text{First-Order Jensen bound}} \overbrace{\lesssim}^{\text{w.p. } (1-\xi)} \underbrace{\mathbb{E}_\rho[\hat{L}(\boldsymbol{\theta}, D)] + \frac{KL(\rho, \pi)}{n} + cte}_{\text{PAC-Bayes bound (Alquier et al. 2016)}}$$

- The Bayesian posterior **converges** to the minimum of $\mathbb{E}_\rho[L(\boldsymbol{\theta})]$.

- The minimum of the first-order Jensen bound $\mathbb{E}_\rho[L(\boldsymbol{\theta})]$ is

# Is the Bayesian approach an optimal learning strategy?

- The **Bayesian learning strategy**,

$$\underbrace{CE(\rho)}_{\text{Generalization Error}} \overbrace{\leq}^{\text{Jensen Inequality}} \underbrace{\mathbb{E}_\rho[L(\boldsymbol{\theta})]}_{\text{First-Order Jensen bound}} \overbrace{\lesssim}^{\text{w.p. } (1-\xi)} \underbrace{\mathbb{E}_\rho[\hat{L}(\boldsymbol{\theta}, D)] + \frac{KL(\rho, \pi)}{n} + cte}_{\text{PAC-Bayes bound (Alquier et al. 2016)}}$$

- The Bayesian posterior **converges** to the minimum of $\mathbb{E}_\rho[L(\boldsymbol{\theta})]$.

- The minimum of the first-order Jensen bound $\mathbb{E}_\rho[L(\boldsymbol{\theta})]$ is

A **Dirac-delta distribution** centered around $\boldsymbol{\theta}^\star_J$

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

# Is the Bayesian approach an optimal learning strategy?

- The **Bayesian learning strategy**,

$$\underbrace{CE(\rho)}_{\text{Generalization Error}} \overbrace{\leq}^{\text{Jensen Inequality}} \underbrace{\mathbb{E}_\rho[L(\boldsymbol{\theta})]}_{\text{First-Order Jensen bound}} \overbrace{\lesssim}^{\text{w.p. } (1-\xi)} \underbrace{\mathbb{E}_\rho[\hat{L}(\boldsymbol{\theta}, D)] + \frac{KL(\rho, \pi)}{n} + cte}_{\text{PAC-Bayes bound (Alquier et al. 2016)}}$$

- The Bayesian posterior **converges** to the minimum of $\mathbb{E}_\rho[L(\boldsymbol{\theta})]$.

- The minimum of the first-order Jensen bound $\mathbb{E}_\rho[L(\boldsymbol{\theta})]$ is

A **Dirac-delta distribution** centered around $\boldsymbol{\theta}^\star_J$

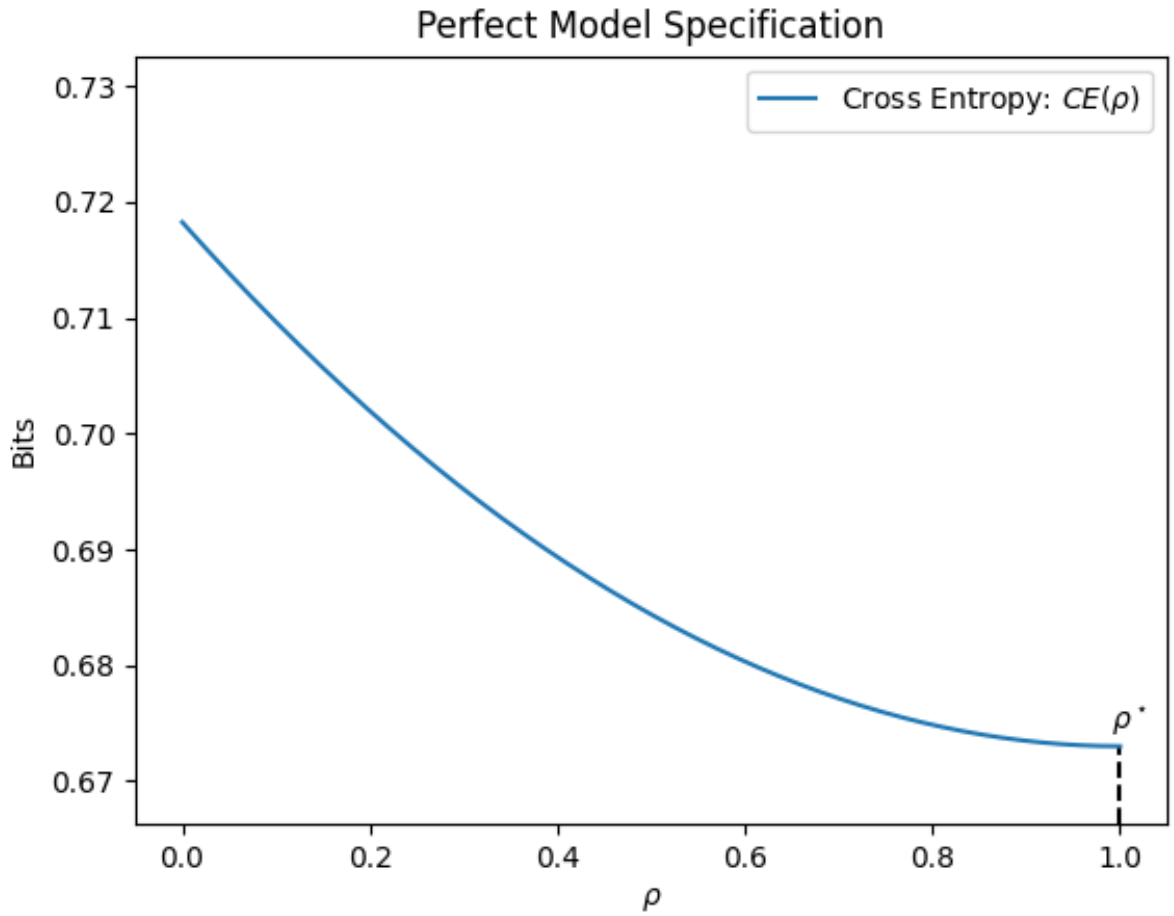
$$\boldsymbol{\theta}^\star_J = \arg\min_\theta KL(\nu(\mathbf{x}), p(\mathbf{x}|\boldsymbol{\theta}))$$

## Is the Bayesian approach an optimal learning strategy?

- Is this **Dirac-delta distribution** centered around $\boldsymbol{\theta}^\star_J$ a good proxy of $\rho^\star$?

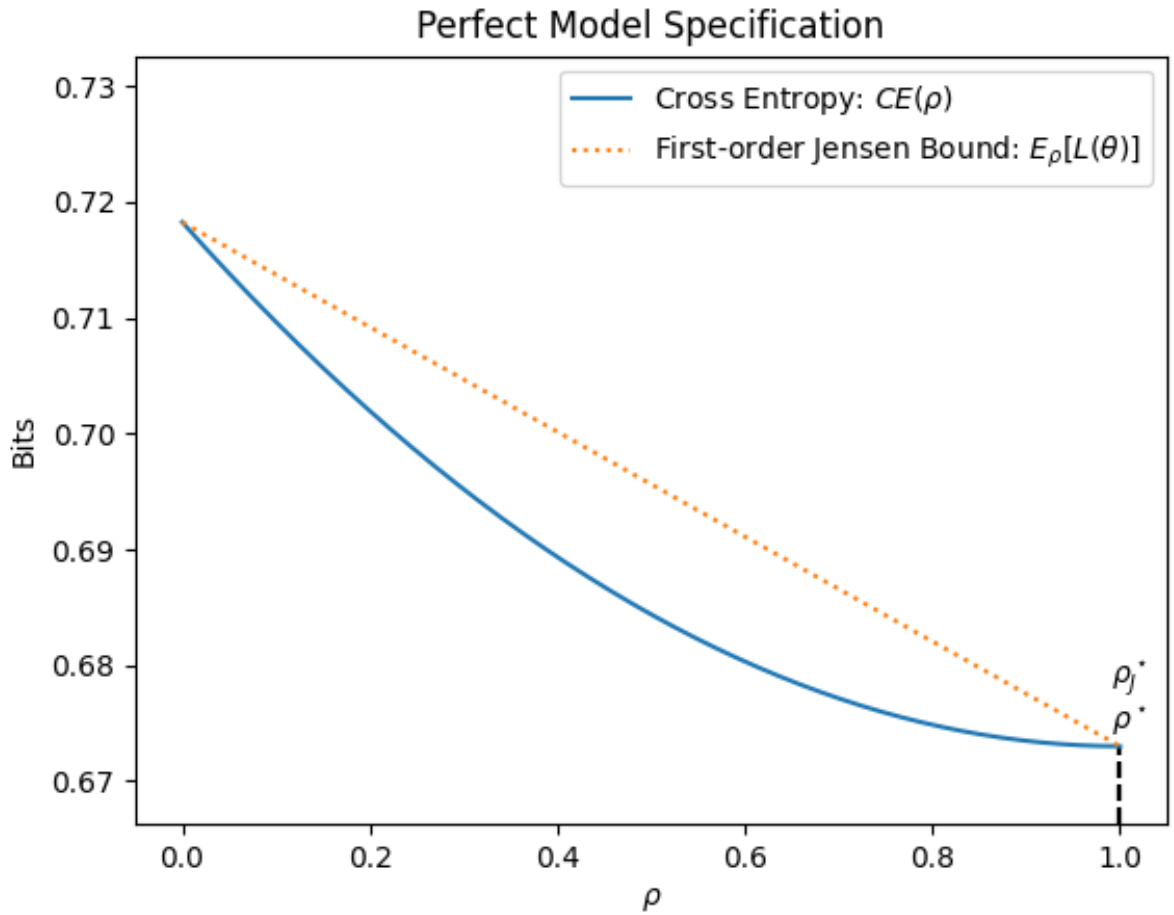
$$\rho^\star = \arg\min_\rho KL(\underbrace{\nu(\mathbf{x})}_{\text{Data distribution}}, \underbrace{\mathbb{E}_{\rho(\theta)}[p(\mathbf{x}|\boldsymbol{\theta})]}_{\text{Predictive posterior}})$$

# Bayesian posterior **optimal under perfect specification**



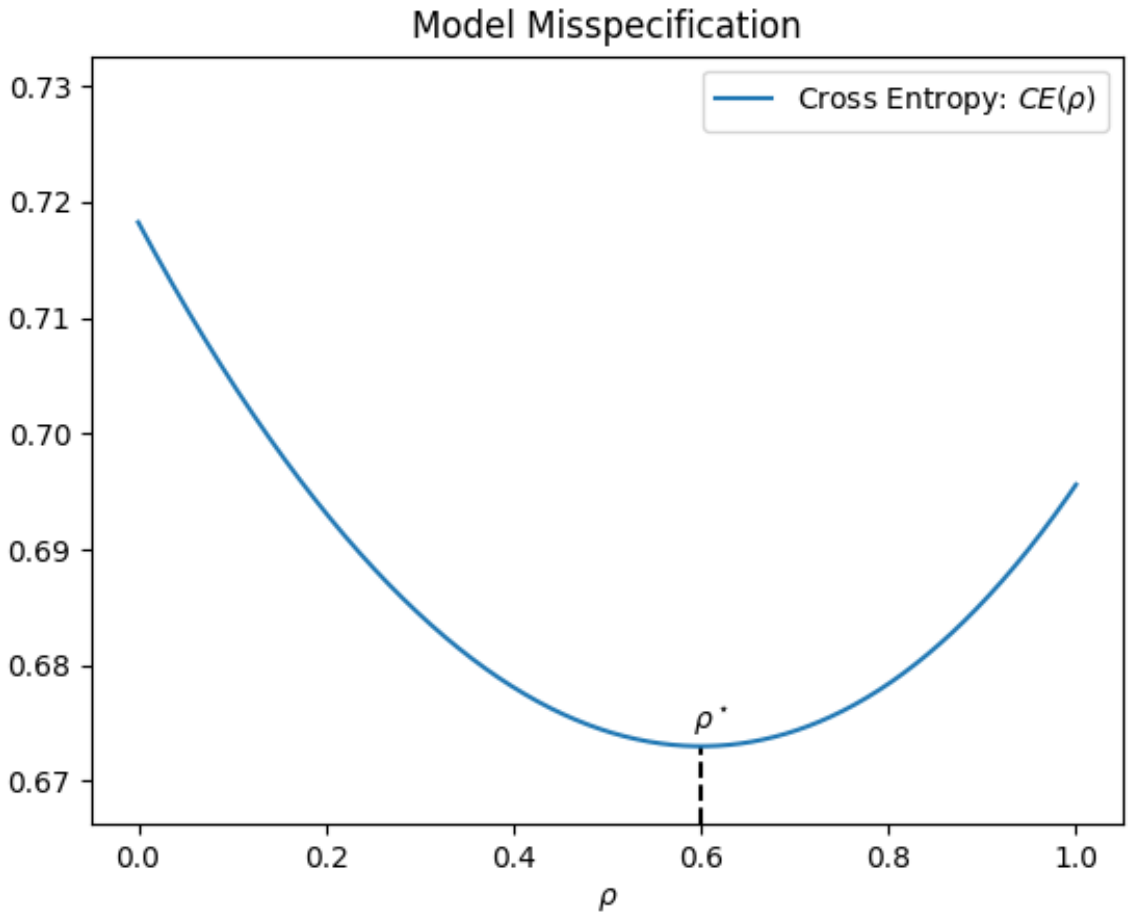

$$\arg\min_{\rho} \underbrace{CE(\rho)}_{\substack{\text{Generalization} \\ \text{Error}}} = \underbrace{\delta_{\theta_J}(\boldsymbol{\theta})}_{\substack{\text{Dirac-Delta} \\ \text{distribution}}}$$

# Bayesian posterior **optimal under perfect specification**



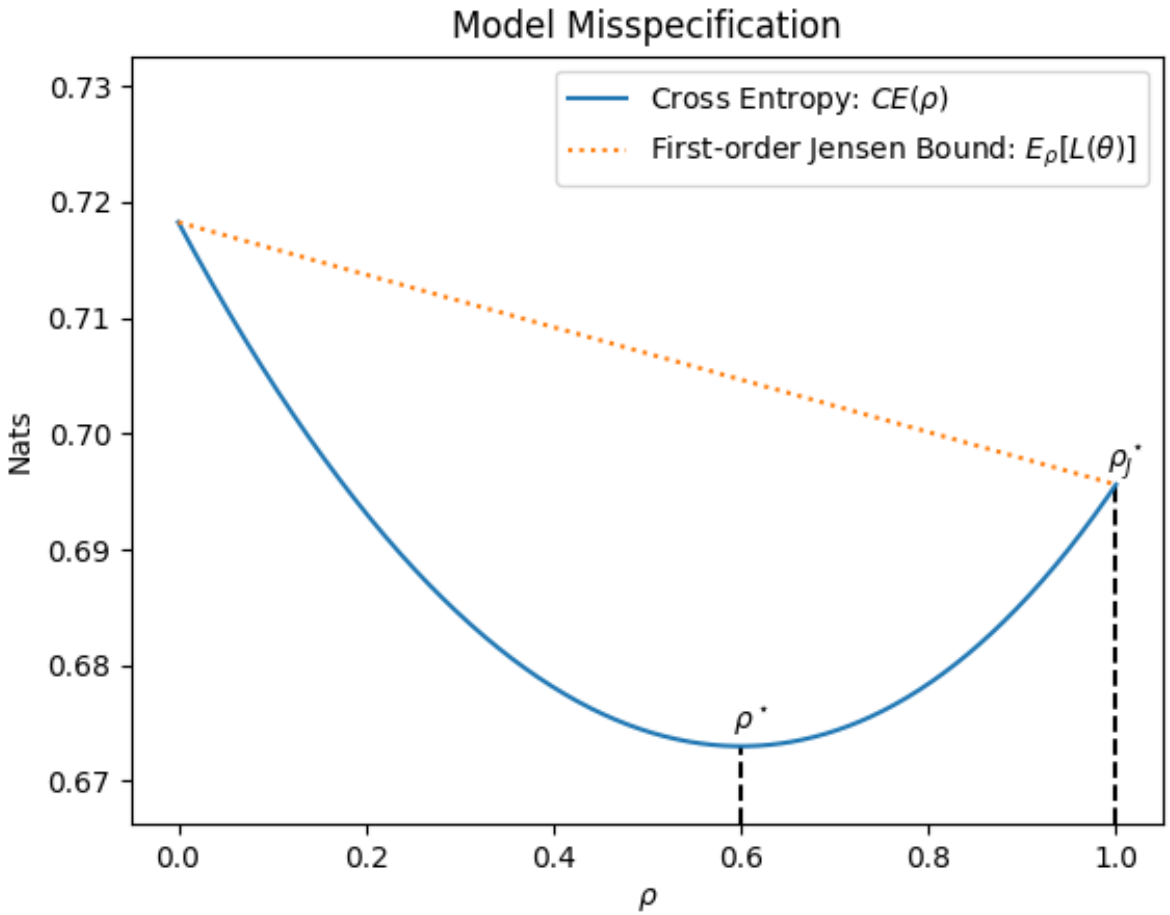

$$\arg\min_{\rho} \underbrace{CE(\rho)}_{\substack{\text{Generalization} \\ \text{Error}}} = \underbrace{\delta_{\theta_J}(\boldsymbol{\theta})}_{\substack{\text{Dirac-Delta} \\ \text{distribution}}} = \arg\min_{\rho} \underbrace{\mathbb{E}_\rho[L(\boldsymbol{\theta})]}_{\substack{\text{First-Order} \\ \text{Jensen Bound}}}$$

# Bayesian posterior **not optimal under misspecification**



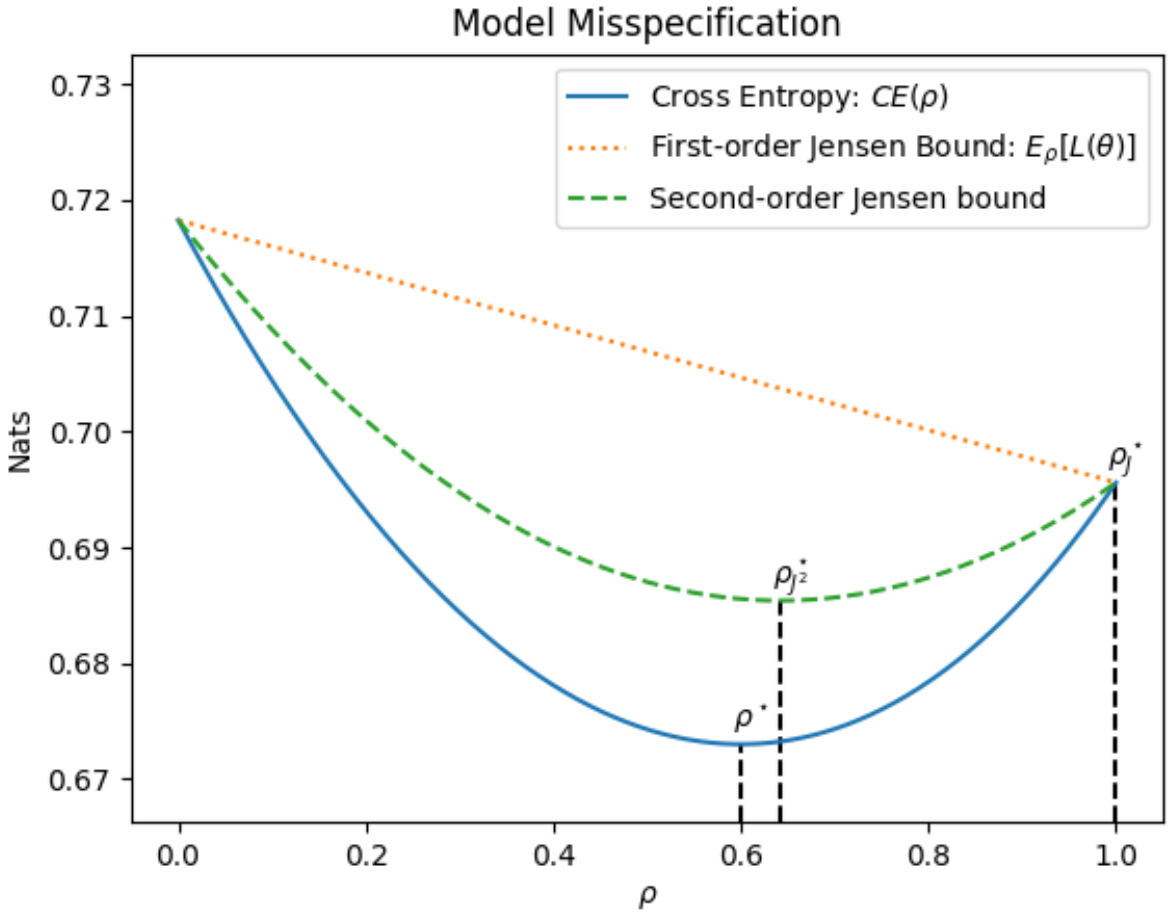

$$\arg\min_{\rho} \underbrace{CE(\rho)}_{\substack{\text{Generalization} \\ \text{Error}}} \neq \underbrace{\delta_{\theta_J}(\boldsymbol{\theta})}_{\substack{\text{Dirac-Delta} \\ \text{distribution}}}$$

# Bayesian posterior **not optimal under misspecification**



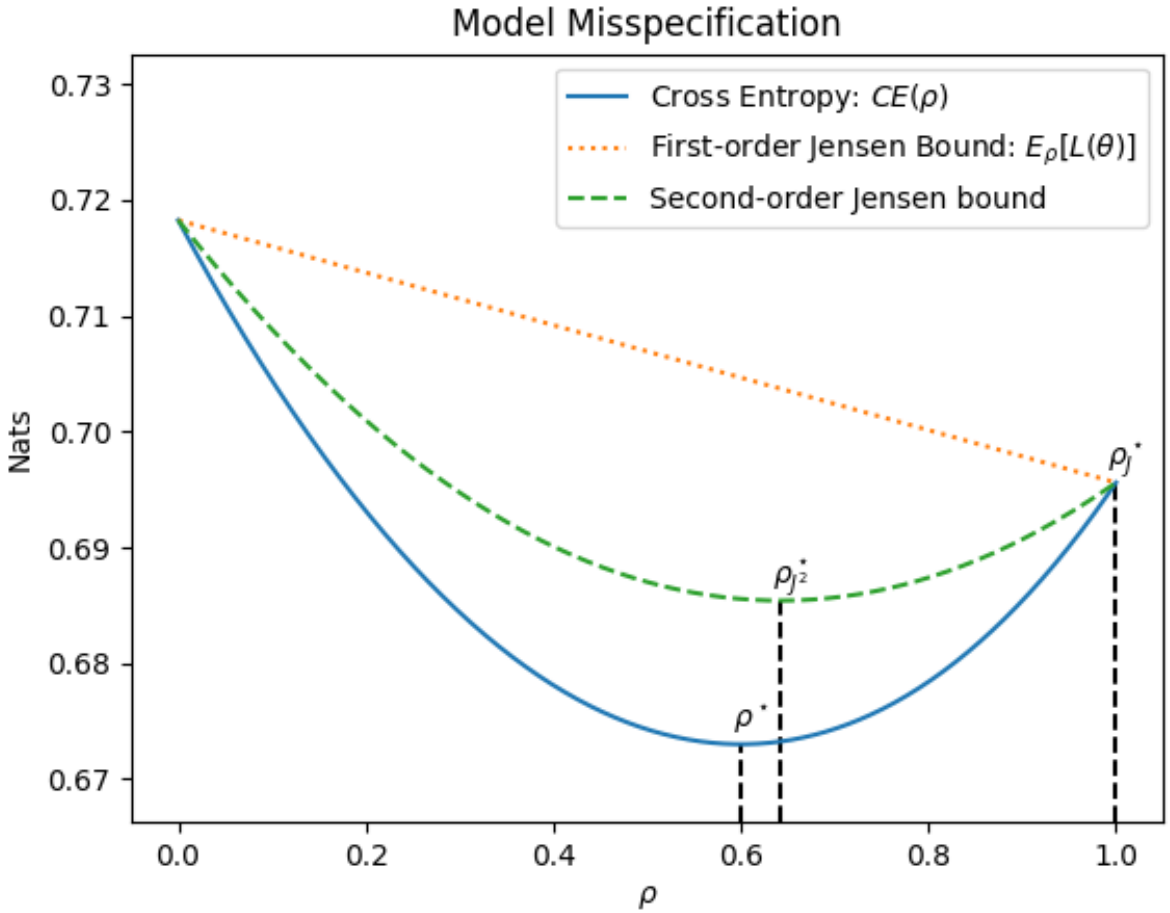

$$\arg\min_{\rho} \underbrace{CE(\rho)}_{\substack{\text{Generalization} \\ \text{Error}}} \neq \underbrace{\delta_{\theta_J}(\boldsymbol{\theta})}_{\substack{\text{Dirac-Delta} \\ \text{distribution}}} = \arg\min_{\rho} \underbrace{\mathbb{E}_\rho[L(\boldsymbol{\theta})]}_{\substack{\text{First-Order} \\ \text{Jensen Bound}}}$$

# Bayesian posterior **not optimal under misspecification**




$$\underbrace{CE(\rho)}_{\text{Generalization Error}} \leq \underbrace{\mathbb{E}_\rho[L(\boldsymbol{\theta})] - \overbrace{\mathbb{V}(\rho)}^{\text{Variance}}}_{\text{Second-order Jensen bound}}$$

# Bayesian posterior **not optimal under misspecification**




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

# Learning by Minimizing second-order PAC-Bayes bounds

## PAC$^2$-Bayesian Learning

- A **variational-like method**,

$$\arg\min_{\rho \in Q} \underbrace{\mathbb{E}_{\rho(\theta)}[L(\boldsymbol{\theta}, D)] - \hat{\mathbb{V}}(\rho, D) + \frac{KL(\rho, \pi)}{n} + cte}_{\text{Second-order PAC-Bayes Bound}}$$

## PAC$^2$-Bayesian Learning

- A **variational-like method**,

$$\arg\min_{\rho \in Q} \underbrace{\mathbb{E}_{\rho(\theta)}[L(\boldsymbol{\theta}, D)] - \hat{\mathbb{V}}(\rho, D) + \frac{KL(\rho, \pi)}{n} + cte}_{\text{Second-order PAC-Bayes Bound}}$$

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

## PAC$^2$-Bayesian Learning

- A **variational-like method**,

$$\arg\min_{\rho \in Q} \underbrace{\mathbb{E}_{\rho(\theta)}[L(\boldsymbol{\theta}, D)] - \hat{\mathbb{V}}(\rho, D) + \frac{KL(\rho, \pi)}{n} + cte}_{\text{Second-order PAC-Bayes Bound}}$$

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

- This is a **generalized variational inference** method (Knoblauch et al. 2019).

## PAC$^2$-Bayesian Learning

- A **variational-like method**,

$$\arg\min_{\rho \in Q} \underbrace{\mathbb{E}_{\rho(\theta)}[L(\boldsymbol{\theta}, D)] - \hat{\mathbb{V}}(\rho, D) + \frac{KL(\rho, \pi)}{n} + cte}_{\text{Second-order PAC-Bayes Bound}}$$

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

- This is a **generalized variational inference** method (Knoblauch et al. 2019).

- Different **solvers** are available in the literature (Wang et al. 2017).

## PAC$^2$-Bayesian Learning

- A **variational-like method**,

$$\arg\min_{\rho \in Q} \underbrace{\mathbb{E}_{\rho(\theta)}[L(\boldsymbol{\theta}, D)] - \hat{\mathbb{V}}(\rho, D) + \frac{KL(\rho, \pi)}{n} + cte}_{\text{Second-order PAC-Bayes Bound}}$$

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

- This is a **generalized variational inference** method (Knoblauch et al. 2019).

- Different **solvers** are available in the literature (Wang et al. 2017).

## Variational Inference

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

$$\arg\min_{\rho \in Q} \underbrace{\mathbb{E}_{\rho(\theta)}[L(\boldsymbol{\theta}, D)] + \frac{KL(\rho, \pi)}{n} + cte}_{\text{First-order PAC-Bayes Bound}}$$

# Ensemble Learning (Finite Mixture Models)

## Mixture of Dirac-delta distribution

- $\rho$ defined as a **mixture of Dirac-delta distributions** parametrized by $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_E\}$,

$$\rho(\boldsymbol{\theta}) = \sum_{j=1}^{E} \frac{1}{E} \delta_{\boldsymbol{\theta}_j}(\boldsymbol{\theta})$$

where $\delta_{\boldsymbol{\theta}_j}$ is a **Dirac-delta distribution** centered around $\boldsymbol{\theta}_j$

## Mixture of Dirac-delta distribution

- $\rho$ defined as a **mixture of Dirac-delta distributions** parametrized by $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_E\}$,

$$\rho(\boldsymbol{\theta}) = \sum_{j=1}^{E} \frac{1}{E} \delta_{\boldsymbol{\theta}_j}(\boldsymbol{\theta})$$

  where $\delta_{\boldsymbol{\theta}_j}$ is a **Dirac-delta distribution** centered around $\boldsymbol{\theta}_j$

- The **predictive posterior** is defined as

$$p_E(\mathbf{x}) = \mathbb{E}_{\rho(\theta)}[p(\mathbf{x}|\boldsymbol{\theta})] = \frac{1}{E} \sum_{j=1}^{E} p(\mathbf{x}|\boldsymbol{\theta}_j)$$

## Mixture of Dirac-delta distribution

- $\rho$ defined as a **mixture of Dirac-delta distributions** parametrized by $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_E\}$,

$$\rho(\boldsymbol{\theta}) = \sum_{j=1}^{E} \frac{1}{E} \delta_{\boldsymbol{\theta}_j}(\boldsymbol{\theta})$$

  where $\delta_{\boldsymbol{\theta}_j}$ is a **Dirac-delta distribution** centered around $\boldsymbol{\theta}_j$

- The **predictive posterior** is defined as

$$p_E(\mathbf{x}) = \mathbb{E}_{\rho(\theta)}[p(\mathbf{x}|\boldsymbol{\theta})] = \frac{1}{E} \sum_{j=1}^{E} p(\mathbf{x}|\boldsymbol{\theta}_j)$$

- The **learning problem**,

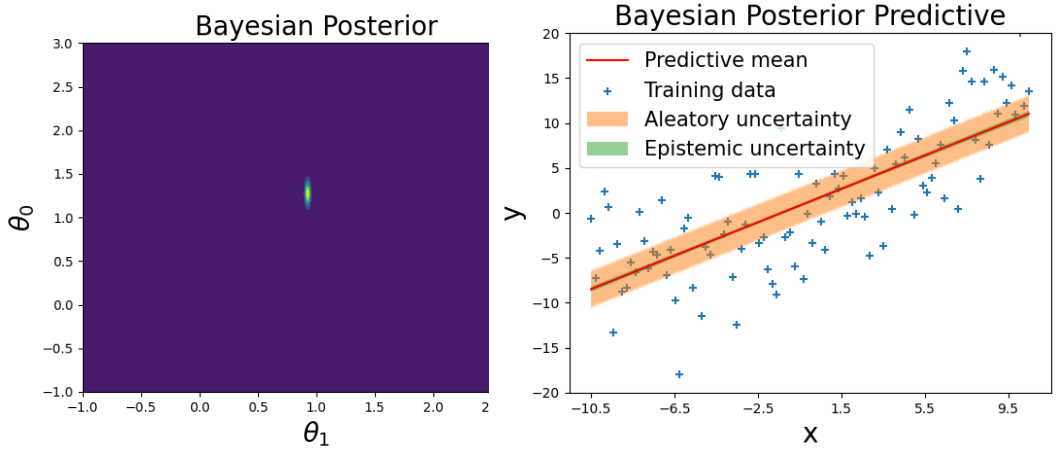
$$\{\boldsymbol{\theta}_1^\star, \ldots, \boldsymbol{\theta}_E^\star\} = \arg \min_{\{\theta_1, \ldots, \theta_E\}} KL(\nu(\mathbf{x}), p_E(\mathbf{x}))$$

# Experimental Evaluation with Toy Data Sets

# Misspecified Linear Regression Model

$$
\begin{aligned}
\nu(y|x) &= \mathcal{N}(\mu = 1 + x, \sigma^2 = 5) \\
p(y|x, \boldsymbol{\theta}) &= \mathcal{N}(\mu = \theta_0 + \theta_1 x, \sigma^2 = 1) \\
\rho(\boldsymbol{\theta}) &= \mathcal{MVN}(\mu, \Sigma)
\end{aligned}
$$

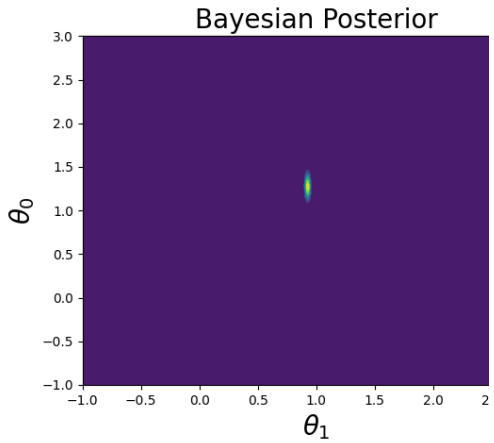
# Misspecified Linear Regression Model

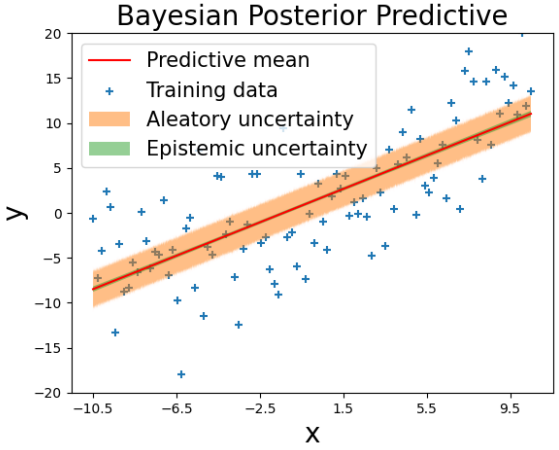



Test Log-likelihood$=-13.09$

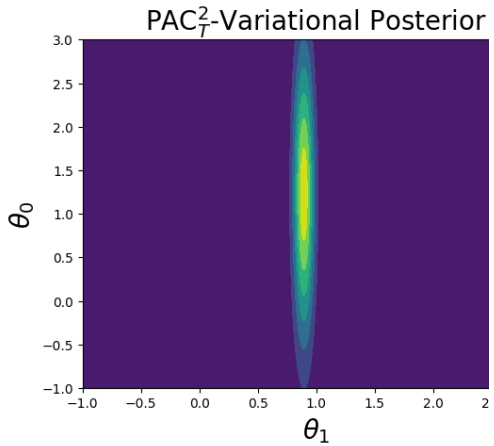
$$
\begin{aligned}
\nu(y|x) &= \mathcal{N}(\mu = 1 + x, \sigma^2 = 5) \\
p(y|x, \boldsymbol{\theta}) &= \mathcal{N}(\mu = \theta_0 + \theta_1 x, \sigma^2 = 1) \\
\rho(\boldsymbol{\theta}) &= \mathcal{MVN}(\mu, \Sigma)
\end{aligned}
$$

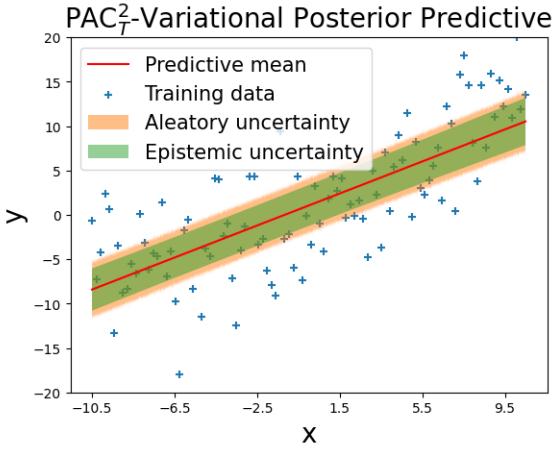
# Misspecified Linear Regression Model



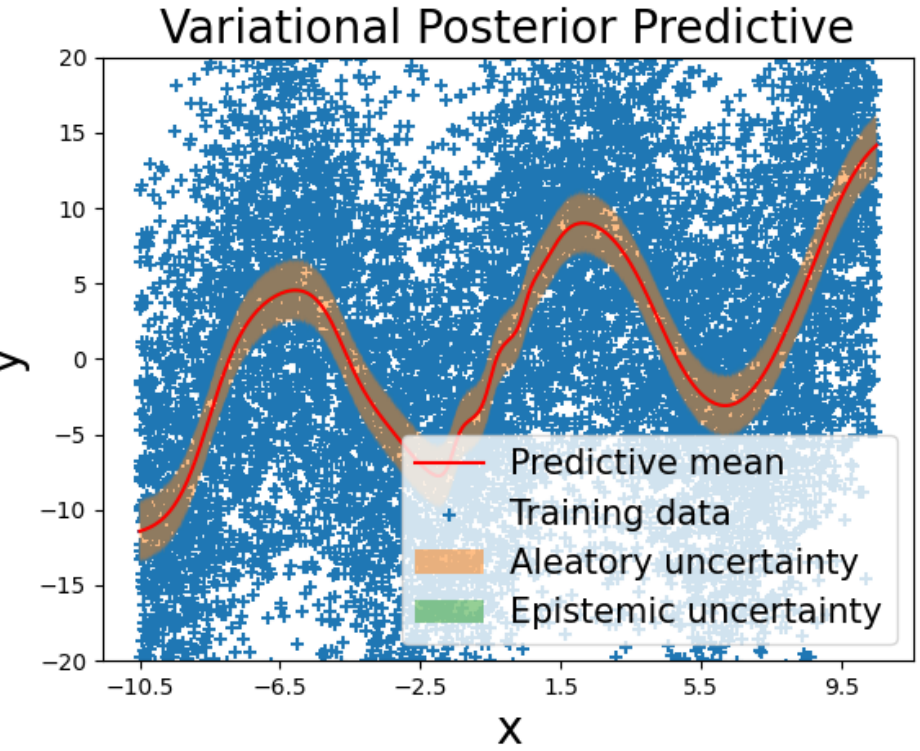

Test Log-likelihood$=-13.09$



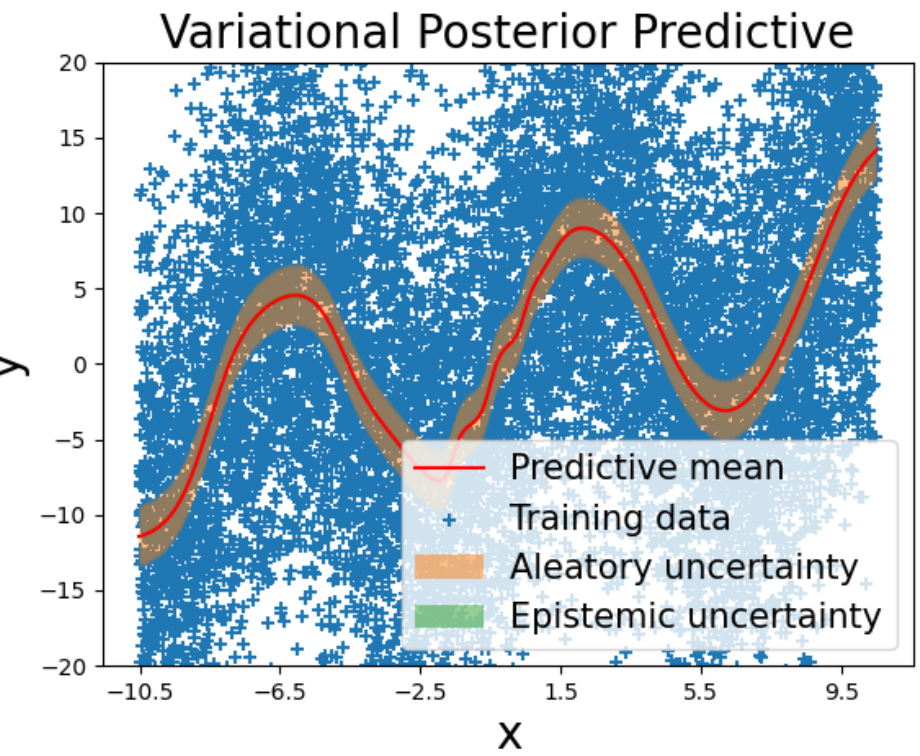

Test Log-likelihood$=-7.89$

$$
\begin{aligned}
\nu(y|x) &= \mathcal{N}(\mu = 1 + x, \sigma^2 = 5) \\
p(y|x, \boldsymbol{\theta}) &= \mathcal{N}(\mu = \theta_0 + \theta_1 x, \sigma^2 = 1) \\
\rho(\boldsymbol{\theta}) &= \mathcal{MVN}(\mu, \Sigma)
\end{aligned}
$$

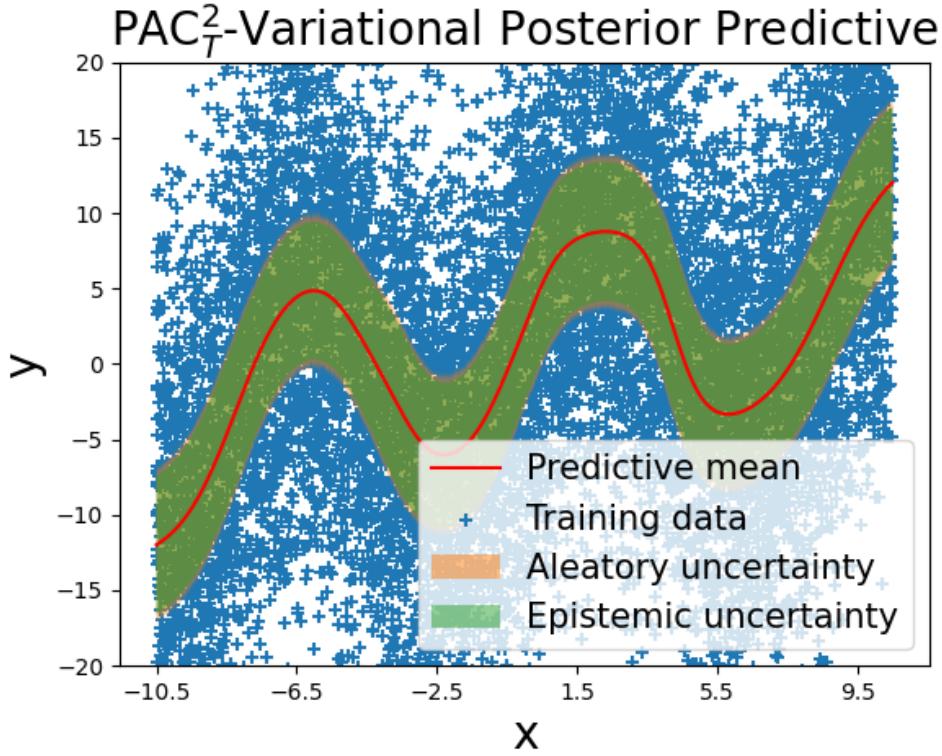



Test Log-likelihood$=-50.15$

$$
\begin{aligned}
\nu(y|x) &= \mathcal{N}(\mu = s(x), \sigma^2 = 10) \\
p(y|x, \boldsymbol{\theta}) &= \mathcal{N}(\mu = MLP_{20}(x; \boldsymbol{\theta}), \sigma^2 = 1) \\
\rho(\boldsymbol{\theta}) &= \prod_i \mathcal{N}(\mu_i, \sigma_i)
\end{aligned}
$$

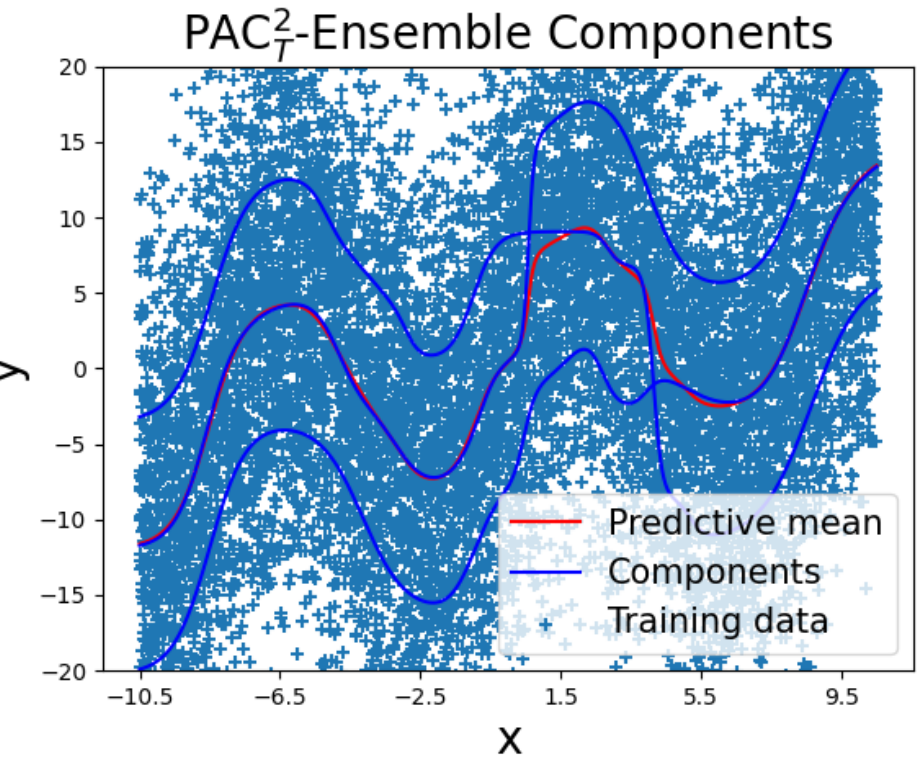


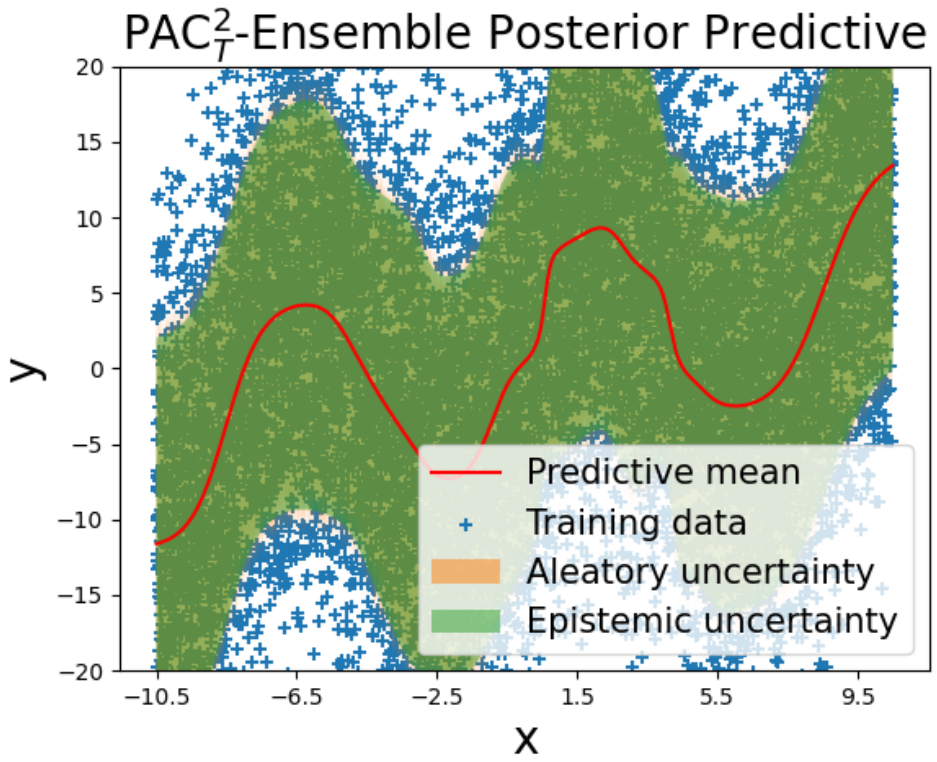

Test Log-likelihood$=-50.15$




Test Log-likelihood$=-25.23$

$$
\begin{aligned}
\nu(y|x) &= \mathcal{N}(\mu = s(x), \sigma^2 = 10) \\
p(y|x, \boldsymbol{\theta}) &= \mathcal{N}(\mu = MLP_{20}(x; \boldsymbol{\theta}), \sigma^2 = 1) \\
\rho(\boldsymbol{\theta}) &= \prod_i \mathcal{N}(\mu_i, \sigma_i)
\end{aligned}
$$



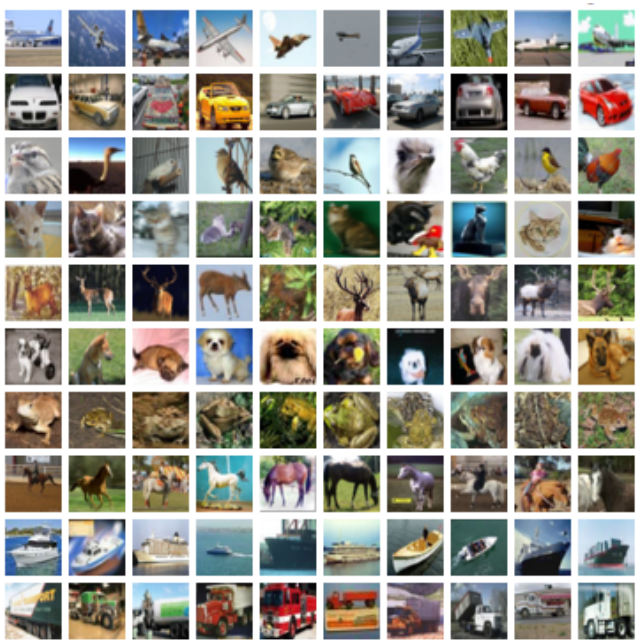

Test Log-likelihood$=-15.91$

$$
\begin{aligned}
\nu(y|x) &= \mathcal{N}(\mu = s(x), \sigma^2 = 10) \\
p(y|x, \boldsymbol{\theta}) &= \mathcal{N}(\mu = MLP_{20}(x; \boldsymbol{\theta}), \sigma^2 = 1) \\
\rho(\boldsymbol{\theta}) &= \sum_{j=1}^{3} \frac{1}{E} \delta_{\boldsymbol{\theta}_j}(\boldsymbol{\theta})
\end{aligned}
$$

# Experimental Evaluation on real data sets

# Data Sets

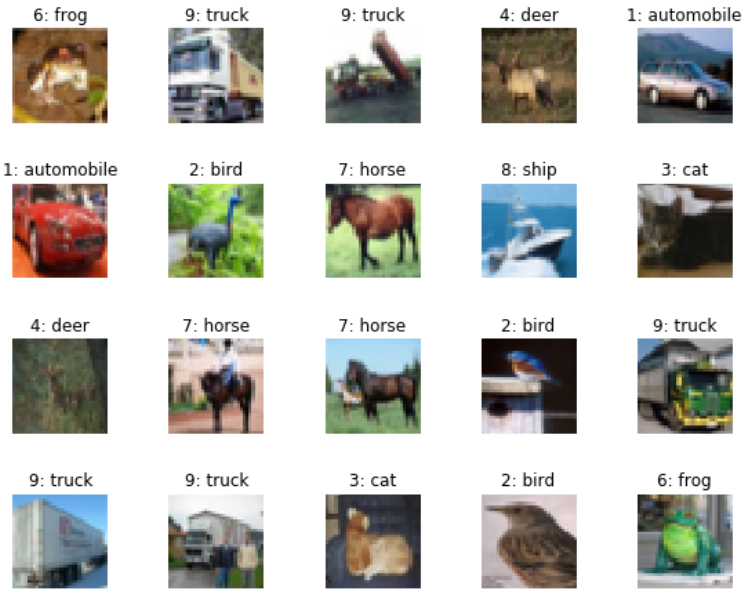
Fahsion-Mnist



CIFAR 10

# Prediction Tasks

Fahsion-Mnist

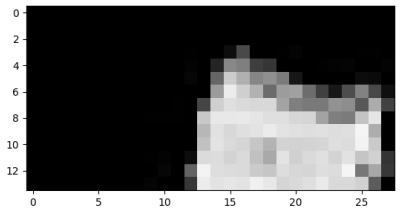



CIFAR 10

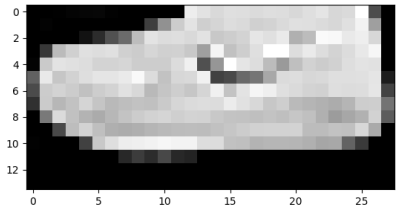
## Task 1

- **Supervised Classification: 10 classes.**

# Prediction Tasks



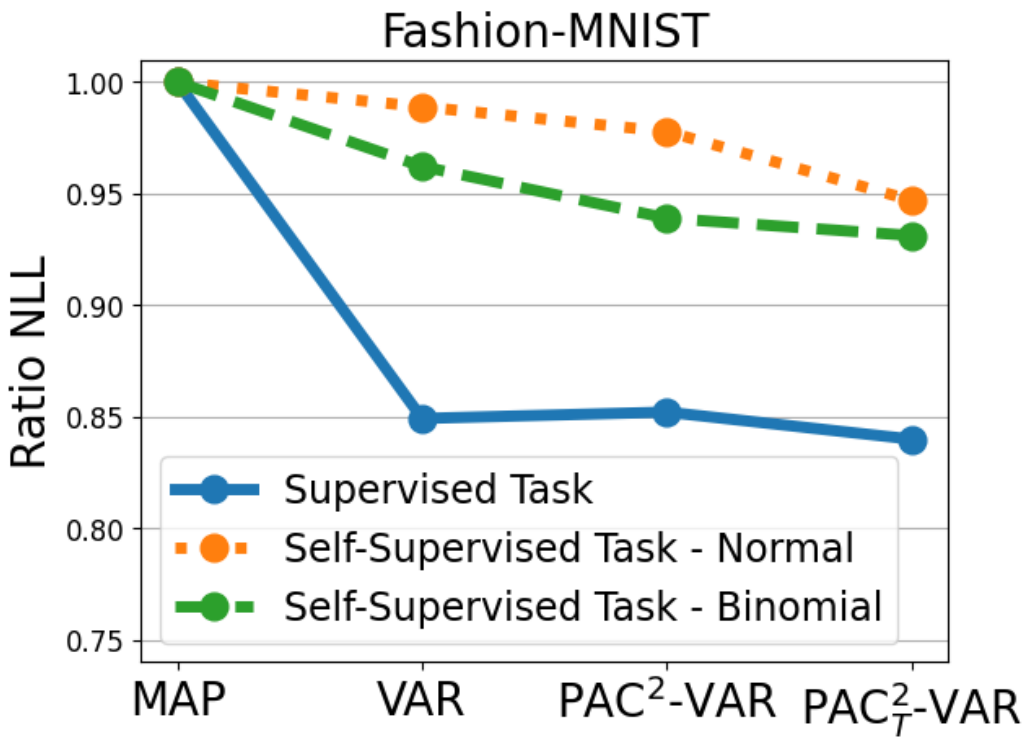

$\mathbf{x}$




$\mathbf{y}$

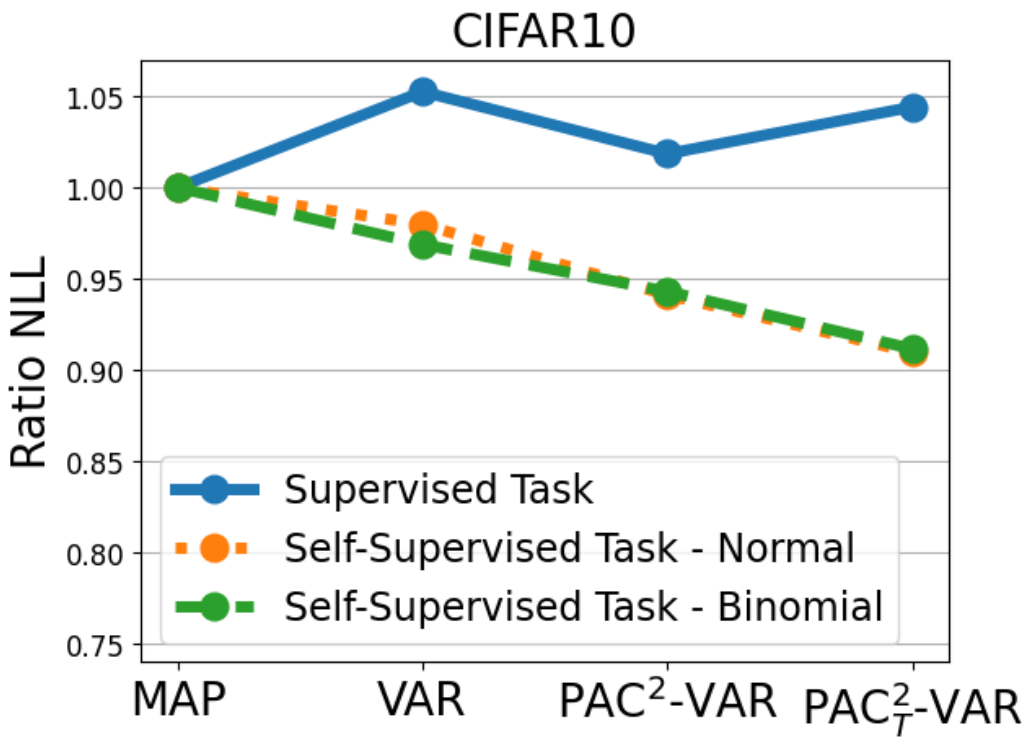
## Self-Supervised Classification

- **Task 2** as a regression/Normal data model.
- **Task 3** as a Binomial data model.




**(the lower the better)**

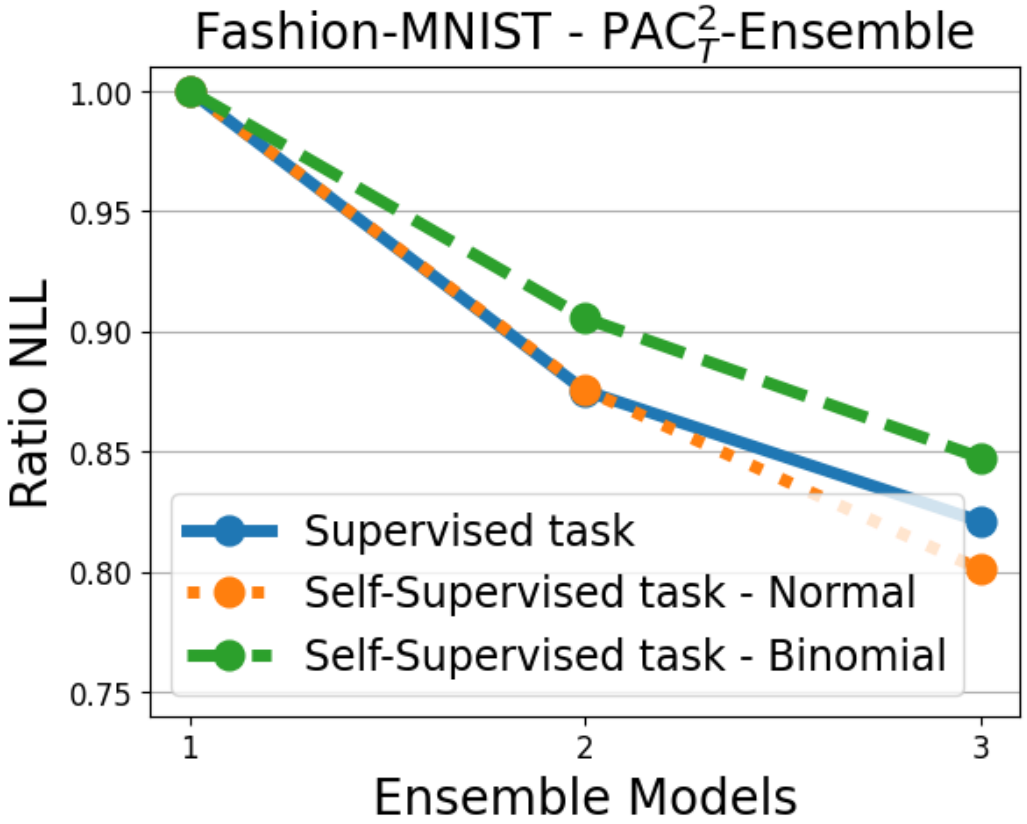
- MLP model with 20 hidden units, Relu activation.
- 100 data batches, 100 epochs, AdamOptimizer default learning rate.




**(the lower the better)**

- MLP model with 20 hidden units, Relu activation.
- 100 data batches, 100 epochs, AdamOptimizer default learning rate.

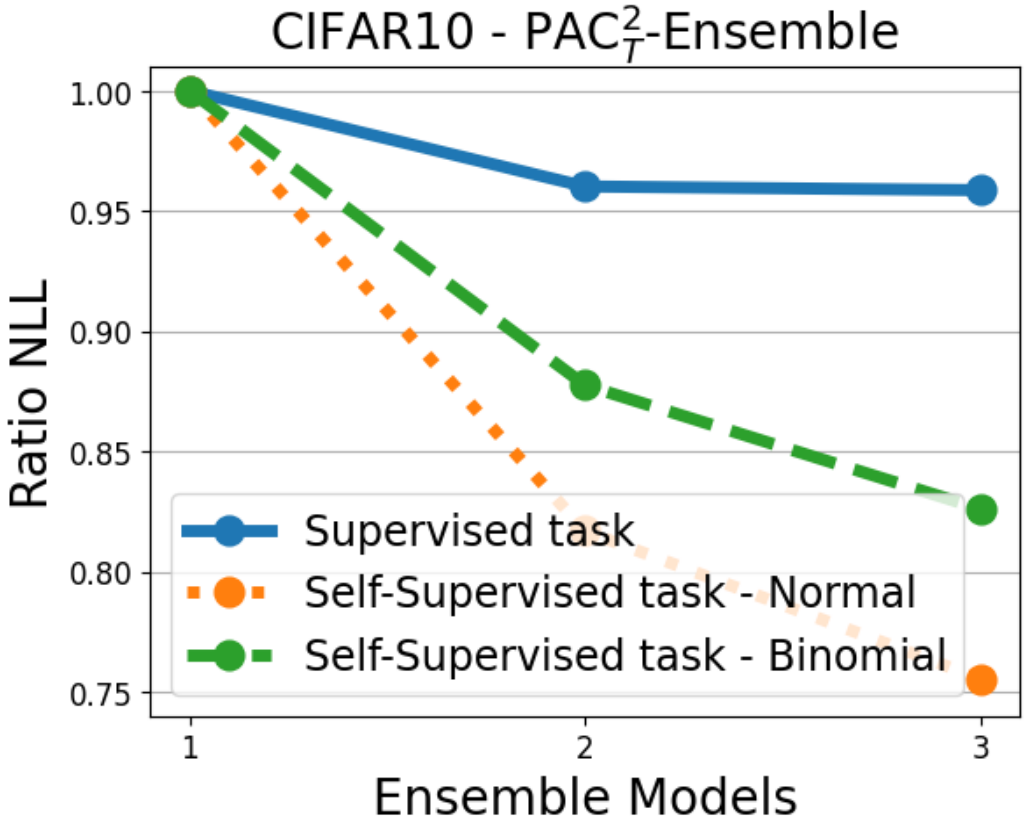
# Ensemble Approach (Finite Mixture Models)




**(the lower the better)**

- **Models initialized with the same parameters**.
- MLP model with 20 hidden units, Relu activation.
- 100 data batches, 100 epochs, AdamOptimizer default learning rate.

# Ensemble Approach (Finite Mixture Models)




**(the lower the better)**

- **Models initialized with the same parameters**.
- MLP model with 20 hidden units, Relu activation.
- 100 data batches, 100 epochs, AdamOptimizer default learning rate.

# Conclusions and Future Works

# Conclusions and Future Works

- The Bayesian approach seems to be **not optimal strategy for learning**.

# Conclusions and Future Works

- The Bayesian approach seems to be **not optimal strategy for learning**.

- **Second-order PAC-Bayesian bounds** directly address mode misspecification.

# Conclusions and Future Works

- The Bayesian approach seems to be **not optimal strategy for learning**.
- **Second-order PAC-Bayesian bounds** directly address mode misspecification.
- Novel **variational and ensemble learning algorithms**.

# Conclusions and Future Works

- The Bayesian approach seems to be **not optimal strategy for learning**.
- **Second-order PAC-Bayesian bounds** directly address mode misspecification.
- Novel **variational and ensemble learning algorithms**.
- **Future works**:
  - Extensive empirical evaluation (new SOTA results in Bayesian deep learning?).

# Conclusions and Future Works

- The Bayesian approach seems to be **not optimal strategy for learning**.
- **Second-order PAC-Bayesian bounds** directly address mode misspecification.
- Novel **variational and ensemble learning algorithms**.
- **Future works**:
  - Extensive empirical evaluation (new SOTA results in Bayesian deep learning?).
  - What happens at the **interpolation regime**?

# Conclusions and Future Works

- The Bayesian approach seems to be **not optimal strategy for learning**.
- **Second-order PAC-Bayesian bounds** directly address mode misspecification.
- Novel **variational and ensemble learning algorithms**.
- **Future works**:
  - Extensive empirical evaluation (new SOTA results in Bayesian deep learning?).
  - What happens at the **interpolation regime**?
  - Related work on Majority Voting:

    Masegosa, A. R., Lorenzen, S. S., Igel, C., & Seldin, Y. Second order PAC-Bayesian bounds for the weighted majority vote. NeurIPS 2020.

# Conclusions and Future Works

- The Bayesian approach seems to be **not optimal strategy for learning**.
- **Second-order PAC-Bayesian bounds** directly address mode misspecification.
- Novel **variational and ensemble learning algorithms**.
- **Future works**:
  - Extensive empirical evaluation (new SOTA results in Bayesian deep learning?).
  - What happens at the **interpolation regime**?
  - Related work on Majority Voting:

    Masegosa, A. R., Lorenzen, S. S., Igel, C., & Seldin, Y. Second order PAC-Bayesian bounds for the weighted majority vote. NeurIPS 2020.

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