Learning under model misspecification

Andrés R. Masegosa

University of Almería
Spain

Masegosa, Andrés R. Learning under model misspecification: Applications to variational and ensemble methods. NeurIPS 2020.

Introduction

Abstract

Model Misspecification

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

Abstract

Model Misspecification

- Virtually any model we use in ML **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**, $D = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$, 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 $\epsilon > 0$),.

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 $\epsilon > 0$),.

This analysis also applies to a supervised settings!!

The Learning Problem

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

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

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

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

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

  - $\mathbb{E}_{\rho(\theta)}[p(\mathbf{x}|\boldsymbol{\theta})]$ is **Bayesian model averaging** when $\rho(\theta) = p(\boldsymbol{\theta}|D)$,

- **Notation**:
  - $\rho(\boldsymbol{\theta})$ is a probability distribution over the parameters of my model.
  - $\rho(\boldsymbol{\theta})$ depends on data. It is a **quasi-posterior**.
- The **predictive posterior distribution** for a given $\rho(\boldsymbol{\theta})$,

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

  - $\mathbb{E}_{\rho(\theta)}[p(\mathbf{x}|\boldsymbol{\theta})]$ is **Bayesian model averaging** when $\rho(\theta) = p(\boldsymbol{\theta}|D)$,
- **The learning problem** is defined as,

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

- **Notation**:
  - $\rho(\boldsymbol{\theta})$ is a probability distribution over the parameters of my model.
  - $\rho(\boldsymbol{\theta})$ depends on data. It is a **quasi-posterior**.
- The **predictive posterior distribution** for a given $\rho(\boldsymbol{\theta})$,

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

  - $\mathbb{E}_{\rho(\theta)}[p(\mathbf{x}|\boldsymbol{\theta})]$ is **Bayesian model averaging** when $\rho(\theta) = p(\boldsymbol{\theta}|D)$,
- **The learning problem** is defined as,

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

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

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

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

  - $\mathbb{E}_{\rho(\theta)}[p(\mathbf{x}|\boldsymbol{\theta})]$ is **Bayesian model averaging** when $\rho(\theta) = p(\boldsymbol{\theta}|D)$,

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

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

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

The learning strategy

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

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

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

The learning strategy

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

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

- 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**:

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

- ... and **minimize** Empirical-Bound(ρ, D, ξ),

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

- 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**:

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

- ... and **minimize** Empirical-Bound(ρ, D, ξ),

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

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

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

$L(\boldsymbol{\theta})$ 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**, $L(\boldsymbol{\theta}, D) = -\frac{1}{n}\ln p(D|\boldsymbol{\theta})$.

$\pi(\boldsymbol{\theta})$ is a prior, which is independent of D.

The Bayesian posterior (Germain et al. 2016)

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

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

The Bayesian posterior (Germain et al. 2016)

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

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

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

The Bayesian posterior (Germain et al. 2016)

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

$$\begin{aligned}\rho^\star &= \arg\min_{\rho} \underbrace{\mathbb{E}_{\rho(\theta)}[L(\boldsymbol{\theta}, D)] + \frac{KL(\rho,\pi)}{n} + \frac{cte}{n}}_{\text{PAC-Bayes bound (Alquier et al. 2016)}} \\ &= \arg\max_{\rho} \underbrace{\mathbb{E}_{\rho(\theta)}[\ln p(D|\boldsymbol{\theta})] - KL(\rho,\pi)}_{\text{Evidence Lower Bound (ELBO)}}\end{aligned}$$

- $\rho^\star$ 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**,

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

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

- $\rho^\star$ 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

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

The **Bayesian learning strategy**,

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

The **Bayesian learning strategy**,

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

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

The **Bayesian learning strategy**,

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

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

The **Bayesian learning strategy**,

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

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

A **Dirac-delta distribution** centered around $\boldsymbol{\theta}^\star_J = \arg\min_\theta KL(\nu(\mathbf{x}), p(\mathbf{x}|\boldsymbol{\theta}))$

The **Bayesian learning strategy**,

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

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

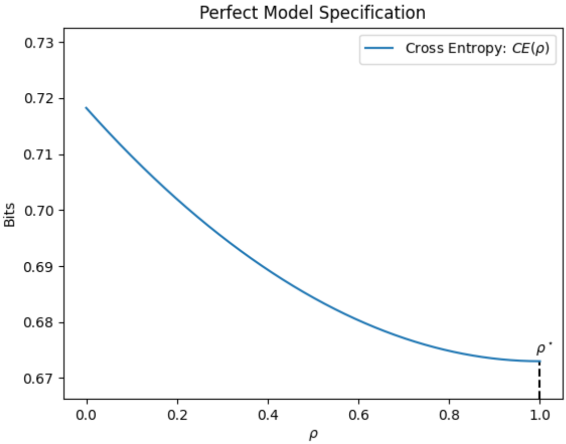
A **Dirac-delta distribution** centered around $\boldsymbol{\theta}^\star_J = \arg\min_\theta KL(\nu(\mathbf{x}), p(\mathbf{x}|\boldsymbol{\theta}))$

Is the Bayesian approach optimal for minimizing the generalization error?

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

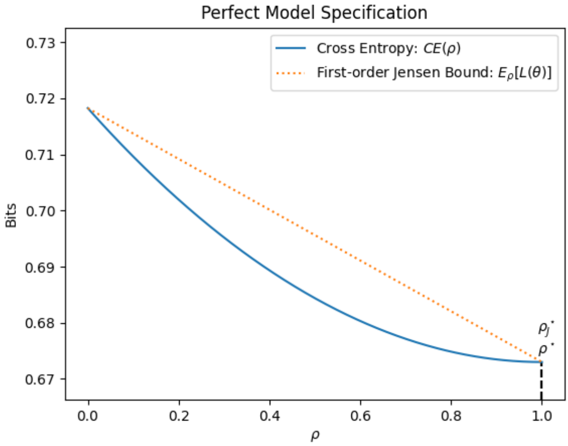
$$\rho^\star = \arg\min_\rho KL(\underbrace{\nu(\mathbf{x})}_{\text{Data distribution}}, \underbrace{\mathbb{E}_{\rho(\theta)}[p(\mathbf{x}|\boldsymbol{\theta})]}_{\text{Predictive posterior}})$$

Bayesian posterior **optimal under perfect specification**



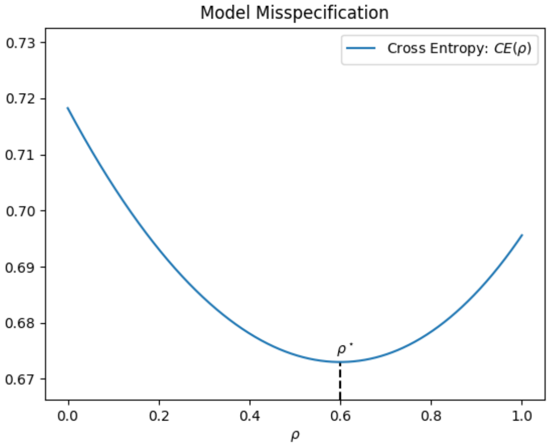
$$\arg\min_{\rho} \underbrace{CE(\rho)}_{\text{Generalization Error}} = \underbrace{\delta_{\theta_J}(\boldsymbol{\theta})}_{\text{Dirac-Delta distribution}}$$

Bayesian posterior **optimal under perfect specification**

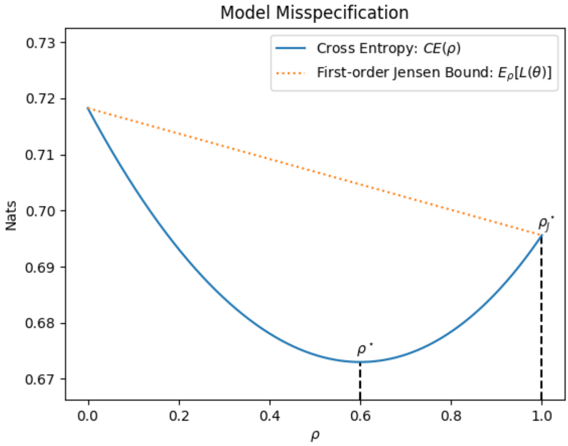


$$\arg\min_\rho \underbrace{CE(\rho)}_{\text{Generalization Error}} = \underbrace{\delta_{\theta_J}(\boldsymbol{\theta})}_{\text{Dirac-Delta distribution}} = \arg\min_\rho \underbrace{\mathbb{E}_\rho[L(\boldsymbol{\theta})]}_{\text{First-Order Jensen Bound}}$$

Bayesian posterior **not optimal under misspecification**

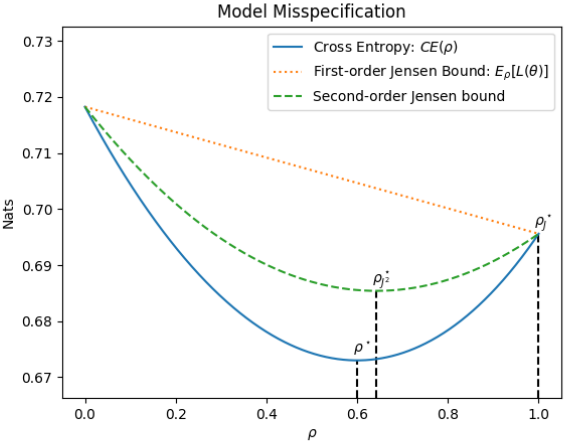


$$\arg\min_{\rho} \underbrace{CE(\rho)}_{\text{Generalization Error}} \neq \underbrace{\delta_{\theta_J}(\boldsymbol{\theta})}_{\text{Dirac-Delta distribution}}$$



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

Bayesian posterior **not optimal under misspecification**



$$\underbrace{CE(\rho)}_{\text{Generalization Error}} \leq \underbrace{\mathbb{E}_\rho[L(\boldsymbol{\theta})] - \overbrace{\mathbb{V}(\rho)}^{\text{Variance}}}_{\text{Second-order Jensen bound (Liao et al. 2019)}}$$

Second-order Jensen Bounds

$$CE(\rho) \leq \underbrace{\mathbb{E}_{\rho}[L(\boldsymbol{\theta})] - \overbrace{\mathbb{V}(\rho)}^{\text{Variance}}}_{\text{Second-order Jensen bound (Liao et al. 2019)}}$$

$$CE(\rho) \leq \underbrace{\mathbb{E}_\rho[L(\boldsymbol{\theta})] - \overbrace{\mathbb{V}(\rho)}^{\text{Variance}}}_{\text{Second-order Jensen bound (Liao et al. 2019)}}$$

- $\mathbb{V}(\rho)$ measures the variance of the predictive posterior:

$$\mathbb{V}(\rho) = \mathbb{E}_{\nu(\mathbf{x})}[\frac{1}{2\max_\theta p(\mathbf{x}|\boldsymbol{\theta})^2}\mathbb{E}_{\rho(\theta)}[(p(\mathbf{x}|\boldsymbol{\theta}) - p(\mathbf{x}))^2]] \geq 0$$

Second-order Jensen Bounds

$$CE(\rho) \leq \underbrace{\mathbb{E}_\rho[L(\boldsymbol{\theta})] - \overbrace{\mathbb{V}(\rho)}^{\text{Variance}}}_{\text{Second-order Jensen bound (Liao et al. 2019)}}$$

- $\mathbb{V}(\rho)$ measures the variance of the predictive posterior:

$$\mathbb{V}(\rho) = \mathbb{E}_{\nu(\mathbf{x})}[\frac{1}{2\max_\theta p(\mathbf{x}|\boldsymbol{\theta})^2}\mathbb{E}_{\rho(\theta)}[(p(\mathbf{x}|\boldsymbol{\theta}) - p(\mathbf{x}))^2]] \geq 0$$

- $\mathbb{V}(\rho)$ accounts for *model diversity*:

$$\mathbb{V}(\rho) = 0 \text{ if } \forall\boldsymbol{\theta} \neq \boldsymbol{\theta}' \;\; p(\mathbf{x}|\boldsymbol{\theta}) = p(\mathbf{x}|\boldsymbol{\theta}')$$

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

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

- $\mathbb{E}_\rho[\hat{L}(\boldsymbol{\theta}, D)]$ encourages to place ρ around individual models with small error.

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

- $\mathbb{E}_\rho[\hat{L}(\boldsymbol{\theta}, D)]$ encourages to place ρ around individual models with small error.
- $\hat{\mathbb{V}}(\rho, D)$ encourages *diversity* among models.

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

- $\mathbb{E}_\rho[\hat{L}(\boldsymbol{\theta}, D)]$ encourages to place ρ around individual models with small error.
- $\hat{\mathbb{V}}(\rho, D)$ encourages *diversity* among models.
  - Key factor when learning under model misspecification.

Second-order PAC-Bayesian Bounds

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

- $\mathbb{E}_\rho[\hat{L}(\boldsymbol{\theta}, D)]$ encourages to place ρ around individual models with small error.
- $\hat{\mathbb{V}}(\rho, D)$ encourages *diversity* among models.
  - Key factor when learning under model misspecification.
- $\frac{KL(\rho,\pi)}{n}$ encourages ρ to be close to π (i.e. acts as a regularizer).

Learning by Minimizing second-order PAC-Bayes bounds

PAC^2-Bayesian Learning

- A **variational-like method**,

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

PAC^2-Bayesian Learning

- A **variational-like method**,

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

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

PAC^2-Bayesian Learning

- A **variational-like method**,

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

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**,

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

  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).

A new learning framework

PAC^2-Bayesian Learning

- A **variational-like method**,

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

  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,

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

Ensembles through Mixture of Dirac-delta distributions

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

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

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

Ensembles through Mixture of Dirac-delta distributions

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

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

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

- **New ensemble learning framework**:

$$\arg\min_{\{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_E\}} \underbrace{\frac{1}{E}\sum_{j=1}^{E} \hat{L}(\boldsymbol{\theta}_j, D)}_{\text{Individual Model Errors}} - \overbrace{\hat{\mathbb{V}}(\rho_E, D)}^{\text{Ensemble Diversity}} - \underbrace{\frac{1}{E}\sum_{j=1}^{E} \frac{\ln \pi(\boldsymbol{\theta}_j)}{n}}_{\text{Regularizer}}$$

Ensembles through Mixture of Dirac-delta distributions

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

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

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

- **New ensemble learning framework**:

$$\arg \min_{\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_E\}} \underbrace{\frac{1}{E} \sum_{j=1}^{E} \hat{L}(\boldsymbol{\theta}_j, D)}_{\text{Individual Model Errors}} - \overbrace{\hat{\mathbb{V}}(\rho_E, D)}^{\text{Ensemble Diversity}} - \underbrace{\frac{1}{E} \sum_{j=1}^{E} \frac{\ln \pi(\boldsymbol{\theta}_j)}{n}}_{\text{Regularizer}}$$

- **Diversity** has been widely recognized as a key factor in ensemble methods:

Ensembles through Mixture of Dirac-delta distributions

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

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

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

- **New ensemble learning framework**:

$$\arg \min_{\{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_E\}} \underbrace{\frac{1}{E}\sum_{j=1}^{E} \hat{L}(\boldsymbol{\theta}_j, D)}_{\text{Individual Model Errors}} - \overbrace{\hat{\mathbb{V}}(\rho_E, D)}^{\text{Ensemble Diversity}} - \underbrace{\frac{1}{E}\sum_{j=1}^{E} \frac{\ln \pi(\boldsymbol{\theta}_j)}{n}}_{\text{Regularizer}}$$

- **Diversity** has been widely recognized as a key factor in ensemble methods:
  - $\mathbb{V}(\rho_E, D)$ is a well-founded diversity measure.

Ensembles through Mixture of Dirac-delta distributions

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

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

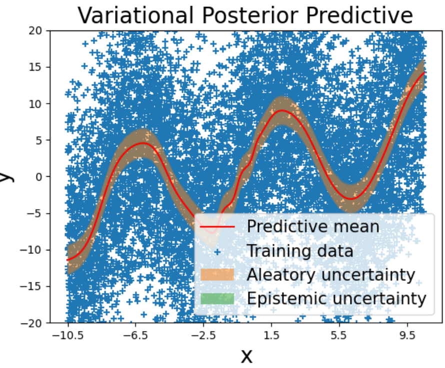
where $\delta_{\boldsymbol{\theta}_j}$ is a **Dirac-delta distribution** centered around $\boldsymbol{\theta}_j$

- **New ensemble learning framework**:

$$\arg \min_{\{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_E\}} \underbrace{\frac{1}{E}\sum_{j=1}^{E} \hat{L}(\boldsymbol{\theta}_j, D)}_{\text{Individual Model Errors}} - \overbrace{\hat{\mathbb{V}}(\rho_E, D)}^{\text{Ensemble Diversity}} - \underbrace{\frac{1}{E}\sum_{j=1}^{E} \frac{\ln \pi(\boldsymbol{\theta}_j)}{n}}_{\text{Regularizer}}$$

- **Diversity** has been widely recognized as a key factor in ensemble methods:
  - $\mathbb{V}(\rho_E, D)$ is a well-founded diversity measure.
  - Help to explain why diversity is key for generalization in ensembles.

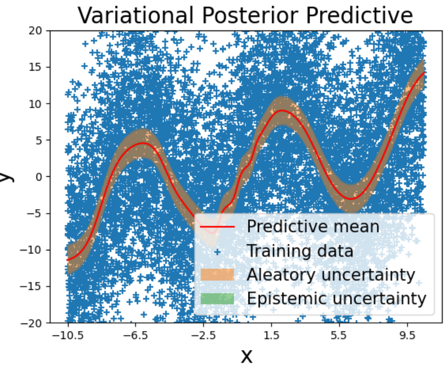
Experimental Evaluation with Toy Data Sets



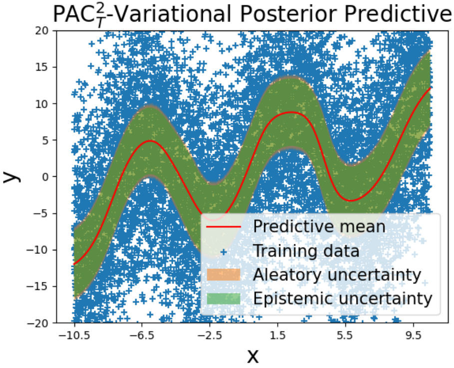
Test Log-likelihood$=-50.15$

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

Bayesian Multilayer Perceptron

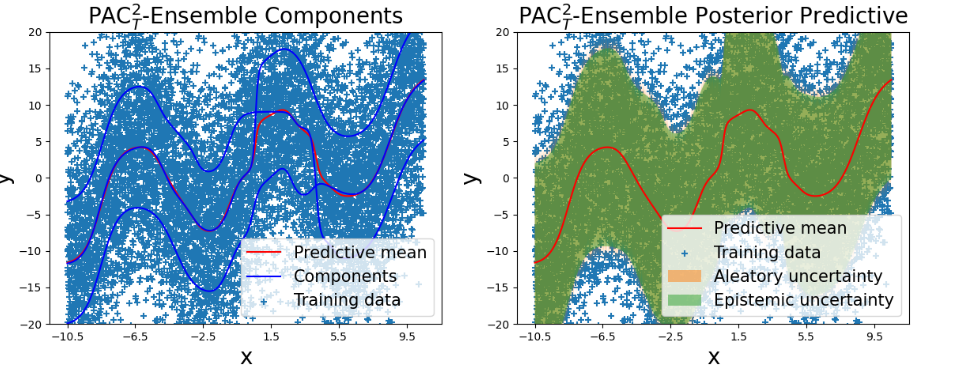


Test Log-likelihood$=-50.15$



Test Log-likelihood$=-25.23$

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



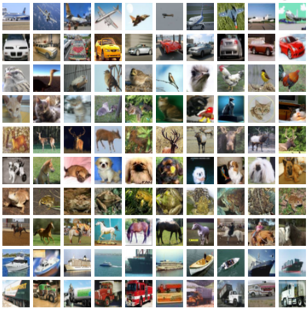
Test Log-likelihood$=-15.91$

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

Experimental Evaluation on real data sets

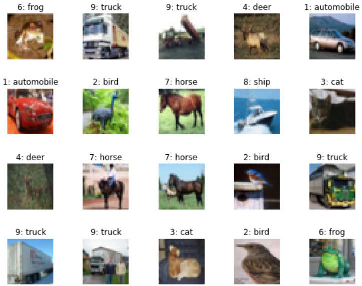
Data Sets

Fahsion-Mnist



CIFAR 10

Fahsion-Mnist

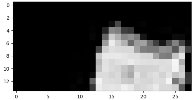


CIFAR 10

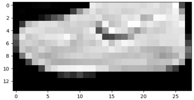
Task 1

- **Supervised Classification: 10 classes.**

Prediction Tasks



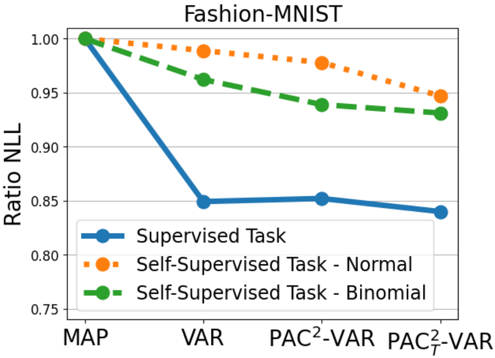
$\mathbf{x}$



$\mathbf{y}$

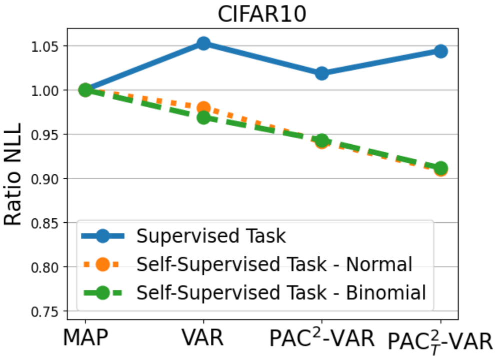
Self-Supervised Classification

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



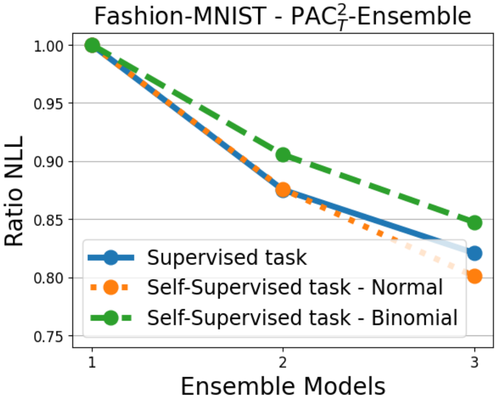
(the lower the better)

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



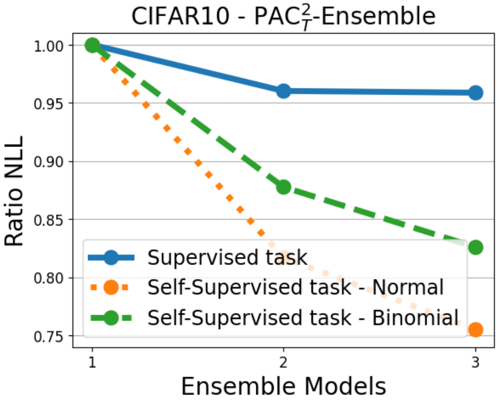
(the lower the better)

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



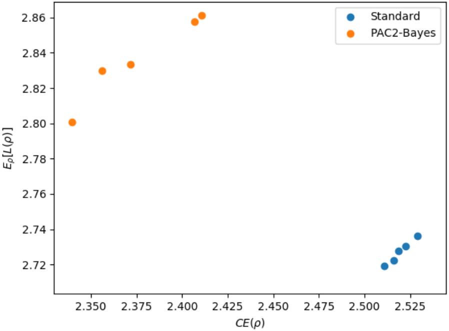
(the lower the better)

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

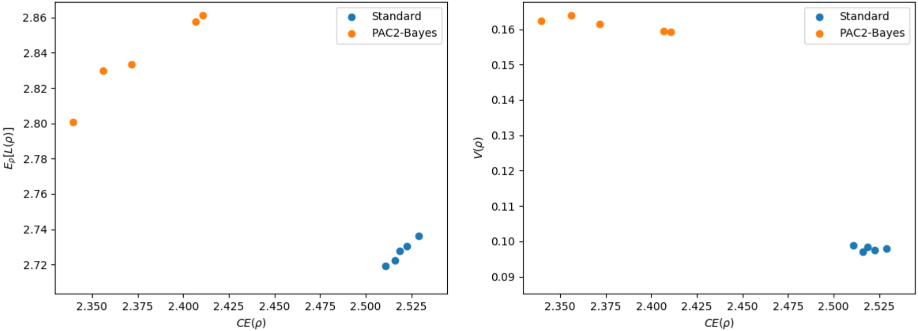


(the lower the better)

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



- Ensembles of four LeNet5 neural network on CIFAR-100



- Ensembles of four LeNet5 neural network on CIFAR-100

Conclusions and Future Works

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

- 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**? In this case, $\mathbb{V}(\rho, D) = 0$

- 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**? In this case, $\mathbb{V}(\rho, D) = 0$
- 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**? In this case, $\mathbb{V}(\rho, D) = 0$
- 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`