On How PAC-Bayesian Bounds Help to Better Understand (and Improve) Bayesian Machine Learning

Andrés R. Masegosa

University of Aalborg (Copenhagen Campus)
Denmark

The Mathematics of Machine Learning Workshop
Bilbao, October, 2022

Bayesian machine learning

The Bayesian approach to Machine Learning

Bayesian machine learning

- Bayesian methods are **widely used** in machine learning.
- They provide well founded approach for dealing with **model/data uncertainty**.
- **Random variables + Probability Calculus**.
- They automatically account for **model complexity**.
- They allow to combine data with **prior knowledge**.

Bayesian Posterior

$$\underbrace{p(\boldsymbol{\theta}|D)}_{\text{Bayesian posterior}} = \frac{\overbrace{p(D|\boldsymbol{\theta})}^{\text{Likelihood}}}{\underbrace{\int p(D|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}_{\text{Normalization Constant}}}\overbrace{\pi(\boldsymbol{\theta})}^{\text{Prior}}$$

Bayesian Posterior

$$\underbrace{p(\boldsymbol{\theta}|D)}_{\text{Bayesian posterior}} = \frac{\overbrace{p(D|\boldsymbol{\theta})}^{\text{Likelihood}}}{\underbrace{\int p(D|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}_{\text{Normalization Constant}}} \overbrace{\pi(\boldsymbol{\theta})}^{\text{Prior}}$$

- Highly multi-modal and complex likelihood for relevant models.
- We have to resort to **approximations** to compute the integral.

Bayesian Posterior

$$\underbrace{p(\boldsymbol{\theta}|D)}_{\text{Bayesian posterior}} = \frac{\overbrace{p(D|\boldsymbol{\theta})}^{\text{Likelihood}}}{\underbrace{\int p(D|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}_{\text{Normalization Constant}}} \overbrace{\pi(\boldsymbol{\theta})}^{\text{Prior}}$$

- Highly multi-modal and complex likelihood for relevant models.
- We have to resort to **approximations** to compute the integral.

Bayesian model averaging

$$\underbrace{p(y_{\text{test}} \mid \mathbf{x}_{\text{test}}, D)}_{\text{Predictive posterior}} = \int \overbrace{p(y_{\text{test}} \mid x_{\text{test}}, \boldsymbol{\theta})}^{\text{Model's prediction}} \underbrace{p(\boldsymbol{\theta}|D)}_{\text{Bayesian posterior}} d\boldsymbol{\theta}$$

Bayesian Posterior

$$\underbrace{p(\boldsymbol{\theta}|D)}_{\text{Bayesian posterior}} = \frac{\overbrace{p(D|\boldsymbol{\theta})}^{\text{Likelihood}}}{\underbrace{\int p(D|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}_{\text{Normalization Constant}}} \overbrace{\pi(\boldsymbol{\theta})}^{\text{Prior}}$$

- Highly multi-modal and complex likelihood for relevant models.
- We have to resort to **approximations** to compute the integral.

Bayesian model averaging

$$\underbrace{p(y_{\text{test}} \mid \mathbf{x}_{\text{test}}, D)}_{\text{Predictive posterior}} = \int \overbrace{p(y_{\text{test}} \mid x_{\text{test}}, \boldsymbol{\theta})}^{\text{Model's prediction}} \quad \underbrace{p(\boldsymbol{\theta}|D)}_{\text{Bayesian posterior}} \quad d\boldsymbol{\theta}$$

- We have to resort to **approximations** to compute the integral.

Bayesian Machine Learning

The main focus of machine learning is the generalization performance
(i.e. how good is my model when making predictions in unseen data samples)

Bayesian Machine Learning

The main focus of machine learning is the generalization performance
(i.e. how good is my model when making predictions in unseen data samples)

Two Main Questions

- Is the **Bayesian posterior** an optimal choice in terms of generalization performance?

The Bayesian approach to Machine Learning

Bayesian Machine Learning

The main focus of machine learning is the generalization performance
(i.e. how good is my model when making predictions in unseen data samples)

Two Main Questions

- Is the **Bayesian posterior** an optimal choice in terms of generalization performance?
- What **kind of priors** should I use to maximize generalization performance?

Is the **Bayesian posterior** optimal for generalization performance?

The learning problem

- $\rho(\boldsymbol{\theta}|D)$ denotes a density over $\boldsymbol{\Theta}$ that depends on the data sample D :
  - $\rho(\boldsymbol{\theta}|D)$ is a **quasi-posterior distribution**.

- $\rho(\boldsymbol{\theta}|D)$ denotes a density over $\boldsymbol{\Theta}$ that depends on the data sample D :
  - $\rho(\boldsymbol{\theta}|D)$ is a **quasi-posterior distribution**.

- Define the **predictive posterior distribution** for a given $\rho(\boldsymbol{\theta}|D)$,

$$\underbrace{p(y_{\text{test}}|x_{\text{test}}, D)}_{\text{Predictive posterior}} = \int \overbrace{p(y_{\text{test}}|x_{\text{test}}, \boldsymbol{\theta})}^{\text{Test likelihood}} \;\underbrace{\rho(\boldsymbol{\theta}|D)}_{\text{Quasi-posterior}}\; d\boldsymbol{\theta}$$

The learning problem

- $\rho(\boldsymbol{\theta}|D)$ denotes a density over $\boldsymbol{\Theta}$ that depends on the data sample D :
  - $\rho(\boldsymbol{\theta}|D)$ is a **quasi-posterior distribution**.

- Define the **predictive posterior distribution** for a given $\rho(\boldsymbol{\theta}|D)$,

$$\underbrace{p(y_{\text{test}}|x_{\text{test}}, D)}_{\text{Predictive posterior}} = \int \overbrace{p(y_{\text{test}}|x_{\text{test}}, \boldsymbol{\theta})}^{\text{Test likelihood}} \ \underbrace{\rho(\boldsymbol{\theta}|D)}_{\text{Quasi-posterior}} \ d\boldsymbol{\theta}$$

  - **Bayesian learning** when $\rho(\boldsymbol{\theta}|D) = p(\boldsymbol{\theta}|D)$,

The learning problem

- $\rho(\boldsymbol{\theta}|D)$ denotes a density over $\boldsymbol{\Theta}$ that depends on the data sample D :
  - $\rho(\boldsymbol{\theta}|D)$ is a **quasi-posterior distribution**.

- Define the **predictive posterior distribution** for a given $\rho(\boldsymbol{\theta}|D)$,

$$\underbrace{p(y_{\text{test}}|x_{\text{test}}, D)}_{\text{Predictive posterior}} = \int \overbrace{p(y_{\text{test}}|x_{\text{test}}, \boldsymbol{\theta})}^{\text{Test likelihood}} \underbrace{\rho(\boldsymbol{\theta}|D)}_{\text{Quasi-posterior}} \; d\boldsymbol{\theta}$$

  - **Bayesian learning** when $\rho(\boldsymbol{\theta}|D) = p(\boldsymbol{\theta}|D)$,

- Generalization performance of $\rho(\boldsymbol{\theta}|D)$:

$$CE(\rho(\theta|D)) = \mathbb{E}_{\nu(\mathbf{y},\mathbf{x})}[-\ln \mathbb{E}_{\rho(\theta|D)}[p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})]]$$

- $\rho(\boldsymbol{\theta}|D)$ denotes a density over $\boldsymbol{\Theta}$ that depends on the data sample D :
  - $\rho(\boldsymbol{\theta}|D)$ is a **quasi-posterior distribution**.

- Define the **predictive posterior distribution** for a given $\rho(\boldsymbol{\theta}|D)$,

$$\underbrace{p(y_{\text{test}}|x_{\text{test}}, D)}_{\text{Predictive posterior}} = \int \overbrace{p(y_{\text{test}}|x_{\text{test}}, \boldsymbol{\theta})}^{\text{Test likelihood}} \; \underbrace{\rho(\boldsymbol{\theta}|D)}_{\text{Quasi-posterior}} \; d\boldsymbol{\theta}$$

  - **Bayesian learning** when $\rho(\boldsymbol{\theta}|D) = p(\boldsymbol{\theta}|D)$,

- Generalization performance of $\rho(\boldsymbol{\theta}|D)$:

$$CE(\rho(\theta|D)) = \mathbb{E}_{\nu(\mathbf{y},\mathbf{x})}[-\ln \mathbb{E}_{\rho(\theta|D)}[p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})]]$$

  - $CE(p(\boldsymbol{\theta}|D))$ measures the **predictive loss of Bayesian learning**.

The learning problem

- In ML, we want to find $\rho(\boldsymbol{\theta}|D)$ which has a small generalization error $CE(\rho(\boldsymbol{\theta}|D))$.

$$\rho^{\star}(\boldsymbol{\theta}|D) = \arg\min_{\rho} CE(\rho(\boldsymbol{\theta}|D))$$

The learning problem

- In ML, we want to find $\rho(\boldsymbol{\theta}|D)$ which has a small generalization error $CE(\rho(\boldsymbol{\theta}|D))$.

$$\rho^\star(\boldsymbol{\theta}|D) = \arg\min_{\rho} CE(\rho(\boldsymbol{\theta}|D))$$

- Is the **Bayesian posterior** the optimal quasi-posterior?

$$p(\boldsymbol{\theta}|D) \approx \rho^\star(\boldsymbol{\theta}|D)$$

PAC-Bayesian Bounds

- Notation:
  - $L(\boldsymbol{\theta})$ is the **expected log-loss**, $L(\boldsymbol{\theta}) = \mathbb{E}_{\nu(\mathbf{y},\mathbf{x})}[-\ln p(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})]$.
  - $\hat{L}(\boldsymbol{\theta}, D)$ is the **empirical log-loss**, $L(\boldsymbol{\theta}, D) = \frac{1}{n}\sum_i -\ln p(\mathbf{y}_i|\mathbf{x}_i,\boldsymbol{\theta})$

- Notation:
  - $L(\boldsymbol{\theta})$ is the **expected log-loss**, $L(\boldsymbol{\theta}) = \mathbb{E}_{\nu(\mathbf{y},\mathbf{x})}[-\ln p(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})]$.
  - $\hat{L}(\boldsymbol{\theta}, D)$ is the **empirical log-loss**, $L(\boldsymbol{\theta}, D) = \frac{1}{n}\sum_i -\ln p(\mathbf{y}_i|\mathbf{x}_i,\boldsymbol{\theta})$

PAC-Bayesian Bound **(Alquier et al. 2016, Germain et al. 2016, Masegosa 2020)**

- For any prior $\pi(\boldsymbol{\theta})$ independent of D and any $\lambda > 0$, for all ρ simultaneously,

$$\underbrace{CE(\rho)}_{\text{Predictive Loss}} \overbrace{\leq}^{\text{Jensen Inequality}} \underbrace{\mathbb{E}_\rho[L(\boldsymbol{\theta})]}_{\text{Gibbs Loss}} \overbrace{\lesssim}^{\text{w.p. } (1-\delta)} \underbrace{\mathbb{E}_\rho[\hat{L}(\boldsymbol{\theta}, D)] + \frac{KL(\rho,\pi)}{\lambda n} + \frac{R_\lambda(\pi)}{\lambda n} + \frac{\ln\frac{1}{\delta}}{\lambda n}}_{\text{PAC-Bayes bound}}$$

- Notation:
  - $L(\boldsymbol{\theta})$ is the **expected log-loss**, $L(\boldsymbol{\theta}) = \mathbb{E}_{\nu(\mathbf{y},\mathbf{x})}[-\ln p(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})]$.
  - $\hat{L}(\boldsymbol{\theta}, D)$ is the **empirical log-loss**, $L(\boldsymbol{\theta}, D) = \frac{1}{n}\sum_i -\ln p(\mathbf{y}_i|\mathbf{x}_i,\boldsymbol{\theta})$

PAC-Bayesian Bound **(Alquier et al. 2016, Germain et al. 2016, Masegosa 2020)**

- For any prior $\pi(\boldsymbol{\theta})$ independent of D and any $\lambda > 0$, for all ρ simultaneously,

$$\underbrace{CE(\rho)}_{\text{Predictive Loss}} \overbrace{\leq}^{\text{Jensen Inequality}} \underbrace{\mathbb{E}_\rho[L(\boldsymbol{\theta})]}_{\text{Gibbs Loss}} \overbrace{\lesssim}^{\text{w.p. } (1-\delta)} \underbrace{\mathbb{E}_\rho[\hat{L}(\boldsymbol{\theta}, D)] + \frac{KL(\rho,\pi)}{\lambda n} + \frac{R_\lambda(\pi)}{\lambda n} + \frac{\ln\frac{1}{\delta}}{\lambda n}}_{\text{PAC-Bayes bound}}$$

  - $R_\lambda(\pi)$ is cummulant-generating function, which is **constant** wrt ρ,

The Bayesian posterior (Germain et al. 2016)

- Which is the quasi-posterior $\rho(\boldsymbol{\theta}|D)$ **minimizing** this PAC-Bayes bound?,

$$\rho^\star(\boldsymbol{\theta}|D) \quad = \quad \arg\min_{\rho} \underbrace{\mathbb{E}_{\rho}[\hat{L}(\boldsymbol{\theta}, D)] + \frac{KL(\rho, \pi)}{\lambda n} + \frac{R_\lambda(\pi)}{\lambda n}}_{\text{PAC-Bayes bound}}$$

The Bayesian posterior (Germain et al. 2016)

- Which is the quasi-posterior $\rho(\boldsymbol{\theta}|D)$ **minimizing** this PAC-Bayes bound?,

$$\rho^\star(\boldsymbol{\theta}|D) \quad = \quad \arg\min_{\rho} \underbrace{\mathbb{E}_\rho[\hat{L}(\boldsymbol{\theta}, D)] + \frac{KL(\rho, \pi)}{\lambda n} + \frac{R_\lambda(\pi)}{\lambda n}}_{\text{PAC-Bayes bound}}$$

- $\rho^\star(\boldsymbol{\theta}|D)$ is the **Bayesian (or Gibbs) posterior**,

$$\rho^\star(\boldsymbol{\theta}|D) = \underbrace{p(\boldsymbol{\theta}|D) \propto p(D|\boldsymbol{\theta})^\lambda \pi(\boldsymbol{\theta})}_{\text{Bayesian posterior}} = \underbrace{e^{-\lambda \hat{L}(\theta, D)} \pi(\boldsymbol{\theta})}_{\text{Gibbs posterior}}$$

The Bayesian posterior (Germain et al. 2016)

- Which is the quasi-posterior $\rho(\boldsymbol{\theta}|D)$ **minimizing** this PAC-Bayes bound?,

$$\rho^\star(\boldsymbol{\theta}|D) \quad = \quad \arg\min_{\rho} \underbrace{\mathbb{E}_{\rho}[\hat{L}(\boldsymbol{\theta}, D)] + \frac{KL(\rho, \pi)}{\lambda n} + \frac{R_\lambda(\pi)}{\lambda n}}_{\text{PAC-Bayes bound}}$$

- $\rho^\star(\boldsymbol{\theta}|D)$ is the **Bayesian (or Gibbs) posterior**,

$$\rho^\star(\boldsymbol{\theta}|D) = \underbrace{p(\boldsymbol{\theta}|D) \propto p(D|\boldsymbol{\theta})^\lambda \pi(\boldsymbol{\theta})}_{\text{Bayesian posterior}} = \underbrace{e^{-\lambda \hat{L}(\theta, D)} \pi(\boldsymbol{\theta})}_{\text{Gibbs posterior}}$$

- Is $p(\boldsymbol{\theta}|D)$ a **good proxy for minimizing** the predictive loss $CE(\rho)$?

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

Key points

- The Bayesian posterior minimizes the PAC-Bayesian upper bound.

PAC-Bayesian bounds and Bayesian Learning (Masegosa, 2020)

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

Key points

- The Bayesian posterior minimizes the PAC-Bayesian upper bound.
- The minimum of the PAC-Bayes bound and $\mathbb{E}_\rho[L(\boldsymbol{\theta})]$ are **aligned**.
  - The Gibbs loss is minimized by a **Dirac-delta** around the best possible model $\boldsymbol{\theta}^\star$

$$\boldsymbol{\theta}^\star = \arg\min_\theta L(\boldsymbol{\theta})$$

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

Key points

- The Bayesian posterior minimizes the PAC-Bayesian upper bound.
- The minimum of the PAC-Bayes bound and $\mathbb{E}_\rho[L(\boldsymbol{\theta})]$ are **aligned**.
  - The Gibbs loss is minimized by a **Dirac-delta** around the best possible model $\boldsymbol{\theta}^\star$

$$\boldsymbol{\theta}^\star = \arg\min_\theta L(\boldsymbol{\theta})$$

  - The Bayesian posterior **converges** to the minimum of the Gibb loss.

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

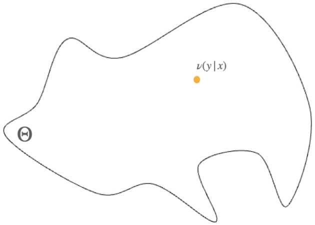
Key points

- The Bayesian posterior minimizes the PAC-Bayesian upper bound.
- The minimum of the PAC-Bayes bound and $\mathbb{E}_\rho[L(\boldsymbol{\theta})]$ are **aligned**.
  - The Gibbs loss is minimized by a **Dirac-delta** around the best possible model $\boldsymbol{\theta}^\star$

$$\boldsymbol{\theta}^\star = \arg\min_\theta L(\boldsymbol{\theta})$$

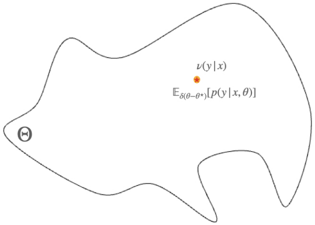
  - The Bayesian posterior **converges** to the minimum of the Gibb loss.
- The minimum of $\mathbb{E}_\rho[L(\boldsymbol{\theta})]$ equals the minimum of $CE(\rho)$ only under **perfect model specification**.

BMA **optimal under perfect specification** (Masegosa, 2020)



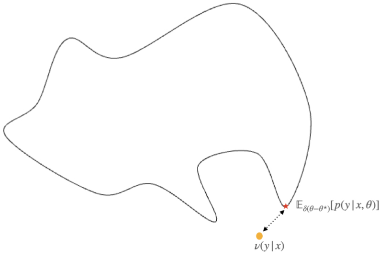
$$\underbrace{\delta(\boldsymbol{\theta} - \boldsymbol{\theta}^{\star})}_{\text{Dirac-Delta}} = \arg\min_{\rho} \underbrace{\mathbb{E}_{\rho}[L(\boldsymbol{\theta})]}_{\textbf{Gibbs Loss}}$$

BMA **optimal under perfect specification** (Masegosa, 2020)



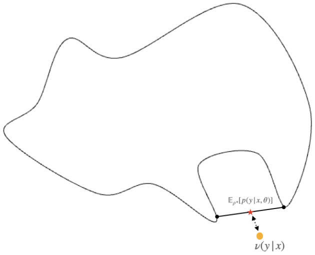
$$\arg\min_{\rho} \underbrace{CE(\rho)}_{\textbf{Predictive Loss}} = \arg\min_{\rho} \underbrace{\mathbb{E}_{\rho}[L(\boldsymbol{\theta})]}_{\textbf{Gibbs Loss}}$$

BMA **not optimal under model misspecification** (Masegosa, 2020)



$$\underbrace{\delta(\boldsymbol{\theta}-\boldsymbol{\theta}^\star)}_{\text{Dirac-Delta}} = \arg\min_{\rho} \underbrace{\mathbb{E}_{\rho}[L(\boldsymbol{\theta})]}_{\textbf{Gibbs Loss}}$$

BMA **not optimal under model misspecification** (Masegosa, 2020)



$$\arg\min_{\rho} \underbrace{CE(\rho)}_{\textbf{Predictive Loss}} \neq \arg\min_{\rho} \underbrace{\mathbb{E}_{\rho}[L(\boldsymbol{\theta})]}_{\textbf{Gibbs Loss}}$$

Predictive PAC-Bayesian Bound (Masegosa, 2020)

For any prior $\pi(\boldsymbol{\theta})$ independent of D and any $\lambda > 0$, for all ρ simultaneously,

$$\underbrace{CE(\rho)}_{\text{Predictive Loss}} \leq \underbrace{\mathbb{E}_{\theta^{(m)}\sim\rho}[L_P(\boldsymbol{\theta}^{(m)})]}_{\text{Gibbs Predictive Loss}} \overbrace{\lesssim}^{\text{w.p. } (1-\delta)} \underbrace{\mathbb{E}_{\theta^{(m)}\sim\rho}[\overbrace{\hat{L}_P(\boldsymbol{\theta}^{(m)}, D)}^{\text{Empirical Predictive Loss}}] + m\frac{KL(\rho,\pi)}{\lambda n} + \frac{R_\lambda(\pi)}{\lambda n}}_{\text{Predictive PAC-Bayes bound}}$$

Predictive PAC-Bayesian Bound (Masegosa, 2020)

For any prior $\pi(\boldsymbol{\theta})$ independent of D and any $\lambda > 0$, for all ρ simultaneously,

$$\underbrace{CE(\rho)}_{\substack{\text{Predictive}\\ \text{Loss}}} \leq \underbrace{\mathbb{E}_{\theta^{(m)}\sim\rho}[L_P(\boldsymbol{\theta}^{(m)})]}_{\substack{\text{Gibbs}\\ \text{Predictive Loss}}} \overbrace{\lesssim}^{\text{w.p. } (1-\delta)} \underbrace{\mathbb{E}_{\theta^{(m)}\sim\rho}[\overbrace{\hat{L}_P(\boldsymbol{\theta}^{(m)}, D)}^{\text{Empirical Predictive Loss}}] + m\frac{KL(\rho,\pi)}{\lambda n} + \frac{R_\lambda(\pi)}{\lambda n}}_{\text{Predictive PAC-Bayes bound}}$$

- $L_P(\boldsymbol{\theta}^{(m)})$ with $m > 1$ induces tighter bounds:

$$\underbrace{CE(\rho)}_{\substack{\text{Predictive}\\ \text{Loss}}} \leq \underbrace{\mathbb{E}_{\theta^{(m)}\sim\rho}[L_P(\boldsymbol{\theta}^{(m)})]}_{\substack{\text{Gibbs}\\ \text{Predictive Loss}}} \leq \underbrace{\mathbb{E}_\rho[L(\boldsymbol{\theta})]}_{\substack{\text{Gibbs}\\ \text{Loss}}}$$

Predictive PAC-Bayesian Bound (Masegosa, 2020)

For any prior $\pi(\boldsymbol{\theta})$ independent of D and any $\lambda > 0$, for all ρ simultaneously,

$$\underbrace{CE(\rho)}_{\substack{\text{Predictive}\\\text{Loss}}} \leq \underbrace{\mathbb{E}_{\theta^{(m)}\sim\rho}[L_P(\boldsymbol{\theta}^{(m)})]}_{\substack{\text{Gibbs}\\\text{Predictive Loss}}} \overset{\text{w.p. } (1-\delta)}{\overbrace{\lesssim}} \underbrace{\mathbb{E}_{\theta^{(m)}\sim\rho}[\overbrace{\hat{L}_P(\boldsymbol{\theta}^{(m)}, D)}^{\text{Empirical Predictive Loss}}] + m\frac{KL(\rho,\pi)}{\lambda n} + \frac{R_\lambda(\pi)}{\lambda n}}_{\text{Predictive PAC-Bayes bound}}$$

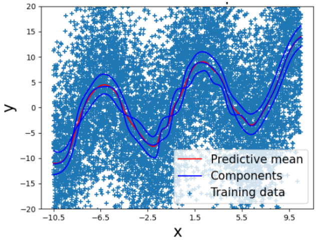
- $L_P(\boldsymbol{\theta}^{(m)})$ with $m > 1$ induces tighter bounds:

$$\underbrace{CE(\rho)}_{\substack{\text{Predictive}\\\text{Loss}}} \leq \underbrace{\mathbb{E}_{\theta^{(m)}\sim\rho}[L_P(\boldsymbol{\theta}^{(m)})]}_{\substack{\text{Gibbs}\\\text{Predictive Loss}}} \leq \underbrace{\mathbb{E}_\rho[L(\boldsymbol{\theta})]}_{\substack{\text{Gibbs}\\\text{Loss}}}$$

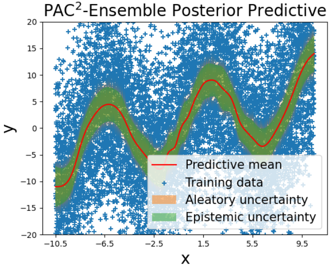
  - (Masegosa, 2020) use a loss based on **second-order Jensen inequalities**.
  - (Futami et al., 2021) use an even **tighter second-order Jensen inequality**.
  - (Morningstar et al., 2022) use a **multi-sample bound** which is **arbitrary tight**.
  - (Zechin et al., 2022) adapts this scheme to **robust** log-losses (t-logarithm).

Experimental Evaluation

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



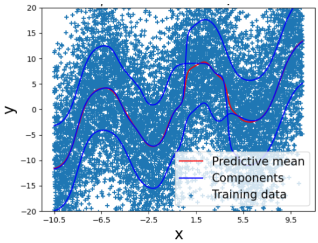
$q(\boldsymbol{\theta}|\gamma)$



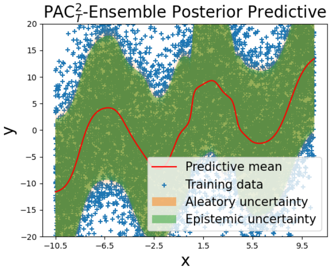
$\mathbb{E}_{q(\theta|\gamma)}[p(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})]$

Standard Bayesian Learning

- Bayesian methods aims to find the **best possible model** within my model class.
- Do not consider the model combination effect.



$q(\boldsymbol{\theta}|\gamma)$



$\mathbb{E}_{q(\theta|\gamma)}[p(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})]$

Predictive Variational Inference

- Aims to find the **best possible model combination**.
- Do consider the **model combination effect**.
- You get that just by **changing the loss function** (i.e. one line of code).

More Experimental Results:
(Masegosa, 2020), (Futami et al., 2021), (Morningstart et al., 2022) and (Zechin et al., 2022).

Bayesian Priors in Bayesian Machine Learning
(work in progress)

Bayesian Priors

Bayesian Priors in Bayesian Statistics

- **(Weakly) Informative Priors**:
  - Priors providing information about the data generating process.
- **(Non-informative) Reference Priors**
  - Priors minimizing the impact they have in the Bayesian posterior.

Bayesian Priors in Bayesian Statistics

- **(Weakly) Informative Priors**:
  - Priors providing information about the data generating process.
- **(Non-informative) Reference Priors**
  - Priors minimizing the impact they have in the Bayesian posterior.

Bayesian Priors in Machine Learning

- **Sparsity-Inducing Priors** (e.g., zero centered Gaussian distributions)
  - Promote small norm parameter that reduce overfitting.
  - Overwhelming empirical evidence.
  - Poorly understood in machine learning.

Bayesian Priors

Bayesian Priors in Bayesian Statistics

- **(Weakly) Informative Priors**:
  - Priors providing information about the data generating process.
- **(Non-informative) Reference Priors**
  - Priors minimizing the impact they have in the Bayesian posterior.

Bayesian Priors in Machine Learning

- **Sparsity-Inducing Priors** (e.g., zero centered Gaussian distributions)
  - Promote small norm parameter that reduce overfitting.
  - Overwhelming empirical evidence.
  - Poorly understood in machine learning.
- What are Sparsity-Inducing Priors?
  - Reference priors, (Weakly) Informative priors or **something different**.

Bayesian Priors

Bayesian Priors in Bayesian Statistics

- **(Weakly) Informative Priors**:
  - Priors providing information about the data generating process.
- **(Non-informative) Reference Priors**
  - Priors minimizing the impact they have in the Bayesian posterior.

Bayesian Priors in Machine Learning

- **Sparsity-Inducing Priors** (e.g., zero centered Gaussian distributions)
  - Promote small norm parameter that reduce overfitting.
  - Overwhelming empirical evidence.
  - Poorly understood in machine learning.
- What are Sparsity-Inducing Priors?
  - Reference priors, (Weakly) Informative priors or **something different**.
- How a **Bayesian prior should look like** to guarantee generalization performance?

PAC-Bayesian Bound

- For any prior $\pi(\boldsymbol{\theta})$ independent of D and any $\lambda > 0$,

$$\underbrace{CE(p_\pi^\lambda)}_{\text{Bayesian Predictive Loss}} \overbrace{\lesssim}^{\text{w.p. } (1-\delta)} \underbrace{-\frac{L\hat{M}_\lambda(\pi, D)}{n\lambda} + \frac{R_\lambda(\pi)}{\lambda n} + \frac{\ln\frac{1}{\delta}}{\lambda n}}_{\text{PAC-Bayes bound}}$$

PAC-Bayesian Bound

- For any prior $\pi(\boldsymbol{\theta})$ independent of D and any $\lambda > 0$,

$$\underbrace{CE(p_\pi^\lambda)}_{\text{Bayesian Predictive Loss}} \overbrace{\lesssim}^{\text{w.p. } (1-\delta)} \underbrace{-\frac{L\hat{M}_\lambda(\pi, D)}{n\lambda} + \frac{R_\lambda(\pi)}{\lambda n} + \frac{\ln\frac{1}{\delta}}{\lambda n}}_{\text{PAC-Bayes bound}}$$

where p_π^λ denotes the **generalized Bayesian posterior**,

$$p_\pi^\lambda(\boldsymbol{\theta}|D) \propto p(D|\boldsymbol{\theta})^\lambda \pi(\boldsymbol{\theta})$$

PAC-Bayesian Bound

- For any prior $\pi(\boldsymbol{\theta})$ independent of D and any $\lambda > 0$,

$$\underbrace{CE(p_\pi^\lambda)}_{\text{Bayesian Predictive Loss}} \overbrace{\lesssim}^{\text{w.p. } (1-\delta)} \underbrace{-\frac{L\hat{M}_\lambda(\pi, D)}{n\lambda} + \frac{R_\lambda(\pi)}{\lambda n} + \frac{\ln\frac{1}{\delta}}{\lambda n}}_{\text{PAC-Bayes bound}}$$

where p_π^λ denotes the **generalized Bayesian posterior**,

$$p_\pi^\lambda(\boldsymbol{\theta}|D) \propto p(D|\boldsymbol{\theta})^\lambda \pi(\boldsymbol{\theta})$$

where $L\hat{M}_\lambda(\pi, D)$ denotes the **log-marginal**:

$$L\hat{M}_\lambda(\pi, D) = \ln \mathbb{E}_\pi[p(D|\boldsymbol{\theta})^\lambda]$$

PAC-Bayesian Bound

- For any prior $\pi(\boldsymbol{\theta})$ independent of D and any $\lambda > 0$,

$$\underbrace{CE(p_\pi^\lambda)}_{\text{Bayesian Predictive Loss}} \quad \overbrace{\lesssim}^{\text{w.p. } (1-\delta)} \quad \underbrace{-\frac{L\hat{M}_\lambda(\pi, D)}{n\lambda} + \frac{R_\lambda(\pi)}{\lambda n} + \frac{\ln\frac{1}{\delta}}{\lambda n}}_{\text{PAC-Bayes bound}}$$

where p_π^λ denotes the **generalized Bayesian posterior**,

$$p_\pi^\lambda(\boldsymbol{\theta}|D) \propto p(D|\boldsymbol{\theta})^\lambda \pi(\boldsymbol{\theta})$$

where $L\hat{M}_\lambda(\pi, D)$ denotes the **log-marginal**:

$$L\hat{M}_\lambda(\pi, D) = \ln \mathbb{E}_\pi[p(D|\boldsymbol{\theta})^\lambda]$$

where $R_\lambda(\pi)$ is a **cummulant generating function**, which can be expressed as:

$$R_\lambda(\pi) = \ln \mathbb{E}_{\pi\nu^n}[e^{\lambda n(L(\theta) - \hat{L}(\theta, D))}]$$

Upper Bounds

- **PAC-Bayesian bound**: For any prior $\pi(\boldsymbol{\theta})$ independent of D and any $\lambda > 0$, ,

$$\underbrace{CE(p_\pi^\lambda)}_{\text{Bayesian Predictive Loss}} \overbrace{\lesssim}^{\text{w.p. } (1-\delta)} \underbrace{-\frac{L\hat{M}_\lambda(\pi, D)}{n\lambda} + \frac{R_\lambda(\pi)}{\lambda n} + \frac{\ln\frac{1}{\delta}}{\lambda n}}_{\text{PAC-Bayes bound}}$$

Upper Bounds

- **PAC-Bayesian bound**: For any prior $\pi(\boldsymbol{\theta})$ independent of D and any $\lambda > 0$, ,

$$\underbrace{CE(p_\pi^\lambda)}_{\text{Bayesian Predictive Loss}} \overbrace{\lesssim}^{\text{w.p. } (1-\delta)} \underbrace{-\frac{L\hat{M}_\lambda(\pi, D)}{n\lambda} + \frac{R_\lambda(\pi)}{\lambda n} + \frac{\ln\frac{1}{\delta}}{\lambda n}}_{\text{PAC-Bayes bound}}$$

- **Expectation bound:** In expectation over different data samples D,

$$\underbrace{\mathbb{E}_D[CE(p_\pi^\lambda)]}_{\text{Bayesian Predictive Loss}} \leq \underbrace{-\frac{\mathbb{E}_D[L\hat{M}_\lambda(\pi, D)]}{n\lambda} + \frac{R_\lambda(\pi)}{\lambda n}}_{\text{Deterministic bound}}$$

Upper Bounds

- **PAC-Bayesian bound**: For any prior $\pi(\boldsymbol{\theta})$ independent of D and any $\lambda > 0$, ,

$$\underbrace{CE(p_\pi^\lambda)}_{\text{Bayesian Predictive Loss}} \overbrace{\lesssim}^{\text{w.p. } (1-\delta)} \underbrace{-\frac{\hat{LM}_\lambda(\pi, D)}{n\lambda} + \frac{R_\lambda(\pi)}{\lambda n} + \frac{\ln\frac{1}{\delta}}{\lambda n}}_{\text{PAC-Bayes bound}}$$

- **Expectation bound:** In expectation over different data samples D,

$$\underbrace{\mathbb{E}_D[CE(p_\pi^\lambda)]}_{\text{Bayesian Predictive Loss}} \leq \underbrace{-\frac{\mathbb{E}_D[\hat{LM}_\lambda(\pi, D)]}{n\lambda} + \frac{R_\lambda(\pi)}{\lambda n}}_{\text{Deterministic bound}}$$

- According to these bounds, **small predictive loss is attained if**:
  - $-\hat{LM}_\lambda(\pi, D)$ **and** $R_\lambda(\pi)$ are both **small**.

Upper Bounds

- **PAC-Bayesian bound**: For any prior $\pi(\boldsymbol{\theta})$ independent of D and any $\lambda > 0$, ,

$$\underbrace{CE(p_\pi^\lambda)}_{\text{Bayesian Predictive Loss}} \overbrace{\lesssim}^{\text{w.p. } (1-\delta)} \underbrace{-\frac{\hat{LM}_\lambda(\pi, D)}{n\lambda} + \frac{R_\lambda(\pi)}{\lambda n} + \frac{\ln\frac{1}{\delta}}{\lambda n}}_{\text{PAC-Bayes bound}}$$

- **Expectation bound:** In expectation over different data samples D,

$$\underbrace{\mathbb{E}_D[CE(p_\pi^\lambda)]}_{\text{Bayesian Predictive Loss}} \leq \underbrace{-\frac{\mathbb{E}_D[\hat{LM}_\lambda(\pi, D)]}{n\lambda} + \frac{R_\lambda(\pi)}{\lambda n}}_{\text{Deterministic bound}}$$

- According to these bounds, **small predictive loss is attained if**:
  - $-\hat{LM}_\lambda(\pi, D)$ **and** $R_\lambda(\pi)$ are both **small**.
  - Both depends on the prior $\pi(\boldsymbol{\theta})$.

Upper Bounds

- **PAC-Bayesian bound**: For any prior $\pi(\boldsymbol{\theta})$ independent of D and any $\lambda > 0$, ,

$$\underbrace{CE(p_\pi^\lambda)}_{\text{Bayesian Predictive Loss}} \overbrace{\lesssim}^{\text{w.p. } (1-\delta)} \underbrace{-\frac{\hat{LM}_\lambda(\pi, D)}{n\lambda} + \frac{R_\lambda(\pi)}{\lambda n} + \frac{\ln\frac{1}{\delta}}{\lambda n}}_{\text{PAC-Bayes bound}}$$

- **Expectation bound:** In expectation over different data samples D,

$$\underbrace{\mathbb{E}_D[CE(p_\pi^\lambda)]}_{\text{Bayesian Predictive Loss}} \leq \underbrace{-\frac{\mathbb{E}_D[\hat{LM}_\lambda(\pi, D)]}{n\lambda} + \frac{R_\lambda(\pi)}{\lambda n}}_{\text{Deterministic bound}}$$

- According to these bounds, **small predictive loss is attained if**:
  - $-\hat{LM}_\lambda(\pi, D)$ **and** $R_\lambda(\pi)$ are both **small**.
  - Both depends on the prior $\pi(\boldsymbol{\theta})$.
  - **Which priors** $\pi(\boldsymbol{\theta})$ make these two terms small?

The Log-Marginal Likelihood

$$L\hat{M}_{\lambda}(\pi, D) = \ln \mathbb{E}_{\pi}[p(D|\boldsymbol{\theta})^{\lambda}]$$

- Widely used in **Bayesian model comparison**.
- Measures how well our model class **explains the data**.
- Depends on the prior $\pi(\boldsymbol{\theta})$.

The Log-Marginal Likelihood

$$L\hat{M}_\lambda(\pi, D) = \ln \mathbb{E}_\pi[p(D|\boldsymbol{\theta})^\lambda]$$

- Widely used in **Bayesian model comparison**.
- Measures how well our model class **explains the data**.
- Depends on the prior $\pi(\boldsymbol{\theta})$.

Theorem: Informative Priors improves the log-marginal likelihood

- Let $\pi_0(\boldsymbol{\theta})$ be a flat or reference prior.
- We build an **informative prior** using (expected) Bayesian updating:

$$\pi_I(\boldsymbol{\theta}) = \mathbb{E}_{D' \sim \nu^n}[p_{\pi_0}^\lambda(\boldsymbol{\theta}|D')]$$

The Log-Marginal Likelihood

$$\hat{LM}_\lambda(\pi, D) = \ln \mathbb{E}_\pi[p(D|\boldsymbol{\theta})^\lambda]$$

- Widely used in **Bayesian model comparison**.
- Measures how well our model class **explains the data**.
- Depends on the prior $\pi(\boldsymbol{\theta})$.

Theorem: Informative Priors improves the log-marginal likelihood

- Let $\pi_0(\boldsymbol{\theta})$ be a flat or reference prior.
- We build an **informative prior** using (expected) Bayesian updating:

$$\pi_I(\boldsymbol{\theta}) = \mathbb{E}_{D' \sim \nu^n}[p^\lambda_{\pi_0}(\boldsymbol{\theta}|D')]$$

- Then, we have that

$$\mathbb{E}_{D\sim\nu^n}[-\hat{LM}_\lambda(\pi_I, D)] \leq \mathbb{E}_{D\sim\nu^n}[-\hat{LM}_\lambda(\pi_0, D)]$$

- Informative priors **reduce**, in expectation, the negative **log-marginal likelihood**.

Upper Bounds

- **PAC-Bayesian bound**: For any prior $\pi(\boldsymbol{\theta})$ independent of D and any $\lambda > 0$, ,

$$\underbrace{CE(p_\pi^\lambda)}_{\text{Bayesian Predictive Loss}} \overbrace{\lesssim}^{\text{w.p. } (1-\delta)} \underbrace{-\frac{L\hat{M}_\lambda(\pi, D)}{n\lambda} + \frac{R_\lambda(\pi)}{\lambda n} + \frac{\ln\frac{1}{\delta}}{\lambda n}}_{\text{PAC-Bayes bound}}$$

- **Expectation bound:** In expectation over different data samples D,

$$\underbrace{\mathbb{E}_D[CE(p_\pi^\lambda)]}_{\text{Bayesian Predictive Loss}} \leq \underbrace{-\frac{\mathbb{E}_D[L\hat{M}_\lambda(\pi, D)]}{n\lambda} + \frac{R_\lambda(\pi)}{\lambda n}}_{\text{Deterministic bound}}$$

Upper Bounds

- **PAC-Bayesian bound**: For any prior $\pi(\boldsymbol{\theta})$ independent of D and any $\lambda > 0$, ,

$$\underbrace{CE(p_\pi^\lambda)}_{\text{Bayesian Predictive Loss}} \overbrace{\lesssim}^{\text{w.p. } (1-\delta)} \underbrace{-\frac{L\hat{M}_\lambda(\pi, D)}{n\lambda} + \frac{R_\lambda(\pi)}{\lambda n} + \frac{\ln\frac{1}{\delta}}{\lambda n}}_{\text{PAC-Bayes bound}}$$

- **Expectation bound:** In expectation over different data samples D,

$$\underbrace{\mathbb{E}_D[CE(p_\pi^\lambda)]}_{\text{Bayesian Predictive Loss}} \leq \underbrace{-\frac{\mathbb{E}_D[L\hat{M}_\lambda(\pi, D)]}{n\lambda} + \frac{R_\lambda(\pi)}{\lambda n}}_{\text{Deterministic bound}}$$

- **Informative priors** reduce the $L\hat{M}_\lambda(\pi, D)$ term:
  - **But not enough** to guarantee generalization performance.
  - Which priors reduce the $R_\lambda(\pi)$ term?

Proposition: $R_\lambda(\pi)$ is a prior regularizer

Over joint draws of $\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta})$ and $D \sim \nu^n(\mathbf{x}, \mathbf{y})$, we have that

$$\underbrace{L(\boldsymbol{\theta}) - \hat{L}(\boldsymbol{\theta}, D)}_{\text{Overfitting}} \overbrace{\lesssim}^{\text{w.p. } (1-\delta)} \frac{1}{\lambda n} R_\lambda(\pi) + \frac{1}{\lambda n} \ln \frac{1}{\delta}. \qquad (1)$$

- If $R_\lambda(\pi)$ is small, then $\pi(\boldsymbol{\theta})$ **prefers models with small overfitting**.

Proposition: $R_\lambda(\pi)$ is a prior regularizer

Over joint draws of $\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta})$ and $D \sim \nu^n(\mathbf{x}, \mathbf{y})$, we have that

$$\underbrace{L(\boldsymbol{\theta}) - \hat{L}(\boldsymbol{\theta}, D)}_{\text{Overfitting}} \overbrace{\lesssim}^{\text{w.p. } (1-\delta)} \frac{1}{\lambda n} R_\lambda(\pi) + \frac{1}{\lambda n} \ln \frac{1}{\delta}. \tag{1}$$

- If $R_\lambda(\pi)$ is small, then $\pi(\boldsymbol{\theta})$ **prefers models with small overfitting**.

Proposition: $R_\lambda(\pi)$ is a prior regularizer

- $R_\lambda(\pi) \geq 0$ for any prior $\pi(\boldsymbol{\theta})$ and any $\lambda \geq 0$.
- $R_\lambda(\pi) = 0$ iif $\pi(\boldsymbol{\theta})$ is Dirac-Delta distribution around $\boldsymbol{\theta}_0$,

$$\underbrace{L(\boldsymbol{\theta}_0) - \hat{L}(\boldsymbol{\theta}_0, D)}_{\text{Overfitting}} = 0$$

  - E.g., A neural network with all the weights set to zero.

The Information-Regularization Trade-off

- **PAC-Bayesian bound**: For any prior $\pi(\boldsymbol{\theta})$ independent of D and any $\lambda > 0$, ,

$$\underbrace{CE(p_\pi^\lambda)}_{\text{Bayesian Predictive Loss}} \overbrace{\lesssim}^{\text{w.p. } (1-\delta)} \underbrace{-\frac{\hat{LM}_\lambda(\pi, D)}{n\lambda} + \frac{R_\lambda(\pi)}{\lambda n} + \frac{\ln\frac{1}{\delta}}{\lambda n}}_{\text{PAC-Bayes bound}}$$

- **Expectation bound:** In expectation over different data samples D,

$$\underbrace{\mathbb{E}_D[CE(p_\pi^\lambda)]}_{\text{Bayesian Predictive Loss}} \leq \underbrace{-\frac{\mathbb{E}_D[\hat{LM}_\lambda(\pi, D)]}{n\lambda} + \frac{R_\lambda(\pi)}{\lambda n}}_{\text{Deterministic bound}}$$

- Priors minimizing these upper-bounds face a **trade-off**:
  - **Informative priors** reduce $\hat{LM}_\lambda(\pi, D)$.
  - **Regularizing priors** reduce $R_\lambda(\pi)$.

The Information-Regularization Trade-off

- **PAC-Bayesian bound**: For any prior $\pi(\boldsymbol{\theta})$ independent of D and any $\lambda > 0$, ,

$$\underbrace{CE(p_\pi^\lambda)}_{\text{Bayesian Predictive Loss}} \overbrace{\lesssim}^{\text{w.p. } (1-\delta)} \underbrace{-\frac{\hat{LM}_\lambda(\pi, D)}{n\lambda} + \frac{R_\lambda(\pi)}{\lambda n} + \frac{\ln\frac{1}{\delta}}{\lambda n}}_{\text{PAC-Bayes bound}}$$

- **Expectation bound:** In expectation over different data samples D,

$$\underbrace{\mathbb{E}_D[CE(p_\pi^\lambda)]}_{\text{Bayesian Predictive Loss}} \leq \underbrace{-\frac{\mathbb{E}_D[\hat{LM}_\lambda(\pi, D)]}{n\lambda} + \frac{R_\lambda(\pi)}{\lambda n}}_{\text{Deterministic bound}}$$

- Priors minimizing these upper-bounds face a **trade-off**:
  - **Informative priors** reduce $\hat{LM}_\lambda(\pi, D)$.
  - **Regularizing priors** reduce $R_\lambda(\pi)$.
- Explains why **log-marginal may not correlate with generalization** (Lotfi et al., 2022).

Theorem: Optimal Priors

- If $\pi_0(\boldsymbol{\theta})$ is a flat or reference prior.
- We define a new priors as:

$$\pi_1(\boldsymbol{\theta}) \propto \underbrace{\mathbb{E}_{D' \sim \nu^n}[p^{\lambda}_{\pi_0}(\boldsymbol{\theta}|D')]}_{\text{Informative Prior}} \; \underbrace{e^{-nJ_\nu(\theta,\lambda)}}_{\text{Regularizing Prior}}$$

where $J_\nu(\theta, \lambda)$ is the so-called **Jensen-Gap function**, defined as:

$$J_\nu(\theta, \lambda) = \ln \mathbb{E}_\nu[p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})] - \mathbb{E}_\nu[\ln p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})]$$

Theorem: Optimal Priors

- If $\pi_0(\boldsymbol{\theta})$ is a flat or reference prior.
- We define a new priors as:

$$\pi_1(\boldsymbol{\theta}) \propto \underbrace{\mathbb{E}_{D' \sim \nu^n}[p^{\lambda}_{\pi_0}(\boldsymbol{\theta}|D')]}_{\text{Informative Prior}} \underbrace{e^{-nJ_\nu(\theta,\lambda)}}_{\text{Regularizing Prior}}$$

where $J_\nu(\theta,\lambda)$ is the so-called **Jensen-Gap function**, defined as:

$$J_\nu(\theta,\lambda) = \ln \mathbb{E}_\nu[p(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})] - \mathbb{E}_\nu[\ln p(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})]$$

- Then, we have that

$$\underbrace{\mathbb{E}_D[CE(p^{\lambda}_{\pi_1})]}_{\text{Bayesian Predictive Loss}} \leq \underbrace{-\frac{\mathbb{E}_D[L\hat{M}_\lambda(\pi_1, D)]}{n\lambda} + \frac{R_\lambda(\pi_1)}{\lambda n}}_{\text{Upper bound for } \pi_1(\boldsymbol{\theta})} \leq \underbrace{-\frac{\mathbb{E}_D[L\hat{M}_\lambda(\pi_0, D)]}{n\lambda} + \frac{R_\lambda(\pi_0)}{\lambda n}}_{\text{Upper bound for } \pi_0(\boldsymbol{\theta})}$$

Regularizing Prior

- We define an **Jensen-Gap prior**:

$$\pi_J(\boldsymbol{\theta}) \propto e^{-nJ_\nu(\theta,\lambda)}$$

- **Naturally emerges** when minimizing a (PAC-Bayes) upper-bound over the Bayesian predictive loss.

- **Proposition:** For any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, over random draws of $D \sim \nu^n(\mathbf{x}, \mathbf{y})$, we have that

$$\underbrace{L(\boldsymbol{\theta}) - \hat{L}(\boldsymbol{\theta}, D)}_{\text{Overfitting}} \overbrace{\lesssim}^{\text{w.p. } (1-\delta)} \frac{1}{\lambda n} J_\nu(\boldsymbol{\theta}, \lambda) + \frac{1}{\lambda n} \ln \frac{1}{\delta}. \tag{2}$$

  - $\pi_R(\boldsymbol{\theta})$ assigns low probability to models with high risk of overfitting.

- $\pi_J(\boldsymbol{\theta})$ *addresses* overfitting (i.e., a regularizing prior).
  - It is a **frequentist prior**.

$\pi_J(\boldsymbol{\theta})$ and existing regularization methods.

MAP estimate using $\pi_J(\boldsymbol{\theta})$

$$\begin{aligned}\theta_{\mathsf{MAP}} &= \arg\max_{\theta} p^{\lambda}_{\pi_J}(\boldsymbol{\theta}|D) \\ &= \arg\min_{\theta} \hat{L}(\boldsymbol{\theta}, D) + \underbrace{\frac{J_\nu(\boldsymbol{\theta}, \lambda)}{\lambda}}_{\text{Regularizer}}\end{aligned}$$

$\pi_J(\boldsymbol{\theta})$ and existing regularization methods.

MAP estimate using $\pi_J(\boldsymbol{\theta})$

$$\begin{aligned}\theta_{\mathsf{MAP}} &= \arg\max_{\theta} p^{\lambda}_{\pi_J}(\boldsymbol{\theta}|D) \\ &= \arg\min_{\theta} \hat{L}(\boldsymbol{\theta}, D) + \underbrace{\frac{J_\nu(\boldsymbol{\theta}, \lambda)}{\lambda}}_{\text{Regularizer}}\end{aligned}$$

$\pi_J(\boldsymbol{\theta})$ and existing regularization methods.

$\pi_J(\boldsymbol{\theta})$ and frequentist estimation theory

Proposition: Under a 2nd-order Taylor approximation of $J_\nu(\boldsymbol{\theta}, \lambda)$ wrt λ:

$$J_\nu(\boldsymbol{\theta}, \lambda) \approx \frac{\lambda^2}{2} \mathbb{V}_{D\sim\nu^n}\left(\hat{L}(\boldsymbol{\theta}, D)\right)$$

- Connection with **frequentist estimation theory**:
  - $\hat{L}(\boldsymbol{\theta}, D)$ is an unbiased estimator of $L(\boldsymbol{\theta})$.
  - $\mathbb{V}_{D\sim\nu^n}\left(\hat{L}(\boldsymbol{\theta}, D)\right)$ is the variance of the estimator.
- Regularization means preferring models with **low variance**.
  - For low variance models, $\hat{L}(\boldsymbol{\theta}, D)$ is a better estimator of $L(\boldsymbol{\theta})$.
- Existing literature: (Namkoong et al. 2017), (Xie et al., 2021), etc.

$\pi_J(\boldsymbol{\theta})$ and L2 regularization (i.e., zero-centered Gaussian priors)

Proposition: For a logistic regression model and under a 2nd-order Taylor approximation of $J_\nu(\boldsymbol{\theta}, \lambda)$ wrt $\boldsymbol{\theta}$:

$$J_\nu(\boldsymbol{\theta}, \lambda) \approx 0.25\lambda^2 \boldsymbol{\theta}^T \mathsf{Cov}_\nu(y\mathbf{x})\boldsymbol{\theta}$$

$\pi_J(\boldsymbol{\theta})$ and existing regularization methods.

$\pi_J(\boldsymbol{\theta})$ and L2 regularization (i.e., zero-centered Gaussian priors)

Proposition: For a logistic regression model and under a 2nd-order Taylor approximation of $J_\nu(\boldsymbol{\theta}, \lambda)$ wrt $\boldsymbol{\theta}$:

$$J_\nu(\boldsymbol{\theta}, \lambda) \approx 0.25\lambda^2 \boldsymbol{\theta}^T \mathsf{Cov}_\nu(y\mathbf{x})\boldsymbol{\theta}$$

- $\pi_J(\boldsymbol{\theta})$ would be a **multivariate normal distribution**:

$$\pi_J(\boldsymbol{\theta}) \propto e^{-n0.25\lambda^2 \theta^T \mathsf{Cov}_\nu(y\mathbf{x})\theta}$$

$\pi_J(\boldsymbol{\theta})$ and existing regularization methods.

$\pi_J(\boldsymbol{\theta})$ and L2 regularization (i.e., zero-centered Gaussian priors)

Proposition: For a logistic regression model and under a 2nd-order Taylor approximation of $J_\nu(\boldsymbol{\theta}, \lambda)$ wrt $\boldsymbol{\theta}$:

$$J_\nu(\boldsymbol{\theta}, \lambda) \approx 0.25\lambda^2 \boldsymbol{\theta}^T \mathsf{Cov}_\nu(y\mathbf{x})\boldsymbol{\theta}$$

- $\pi_J(\boldsymbol{\theta})$ would be a **multivariate normal distribution**:

$$\pi_J(\boldsymbol{\theta}) \propto e^{-n 0.25\lambda^2 \theta^T \mathsf{Cov}_\nu(y\mathbf{x})\theta}$$

- If the data is normalized and features are conditionally independent, it is **equal to L2-regularization** ,

$$\boldsymbol{\theta}^T \mathsf{Cov}_\nu(y\mathbf{x})\boldsymbol{\theta} = \boldsymbol{\theta}^T kI\boldsymbol{\theta} = k||\boldsymbol{\theta}||^2$$

$\pi_J(\boldsymbol{\theta})$ and existing regularization methods.

$\pi_J(\boldsymbol{\theta})$ and L2 regularization (i.e., zero-centered Gaussian priors)

Proposition: For a logistic regression model and under a 2nd-order Taylor approximation of $J_\nu(\boldsymbol{\theta}, \lambda)$ wrt $\boldsymbol{\theta}$:

$$J_\nu(\boldsymbol{\theta}, \lambda) \approx 0.25\lambda^2 \boldsymbol{\theta}^T \mathsf{Cov}_\nu(y\mathbf{x})\boldsymbol{\theta}$$

- $\pi_J(\boldsymbol{\theta})$ would be a **multivariate normal distribution**:

$$\pi_J(\boldsymbol{\theta}) \propto e^{-n0.25\lambda^2 \theta^T \mathsf{Cov}_\nu(y\mathbf{x})\theta}$$

- If the data is normalized and features are conditionally independent, it is **equal to L2-regularization** ,

$$\boldsymbol{\theta}^T \mathsf{Cov}_\nu(y\mathbf{x})\boldsymbol{\theta} = \boldsymbol{\theta}^T kI\boldsymbol{\theta} = k||\boldsymbol{\theta}||^2$$

- **Explains** why L2-regularization improves generalization:
  - Small-norm models tends to have **lower variance**.
  - Lower variance implies **better estimators** $\hat{L}(D, \boldsymbol{\theta})$.
  - Better estimators leads to **less overfitting**.

$\pi_J(\boldsymbol{\theta})$ and existing regularization methods.

$\pi_J(\boldsymbol{\theta})$ and L2 regularization (i.e., zero-centered Gaussian priors)

Proposition: For a logistic regression model and under a 2nd-order Taylor approximation of $J_\nu(\boldsymbol{\theta}, \lambda)$ wrt $\boldsymbol{\theta}$:

$$J_\nu(\boldsymbol{\theta}, \lambda) \approx 0.25\lambda^2 \boldsymbol{\theta}^T \mathsf{Cov}_\nu(y\mathbf{x})\boldsymbol{\theta}$$

- $\pi_J(\boldsymbol{\theta})$ would be a **multivariate normal distribution**:

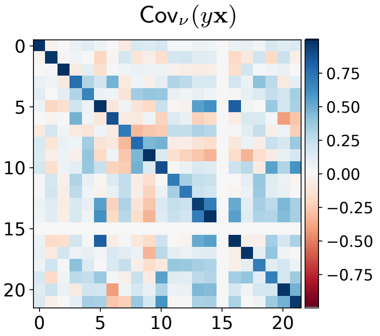
$$\pi_J(\boldsymbol{\theta}) \propto e^{-n0.25\lambda^2 \theta^T \mathsf{Cov}_\nu(y\mathbf{x})\theta}$$

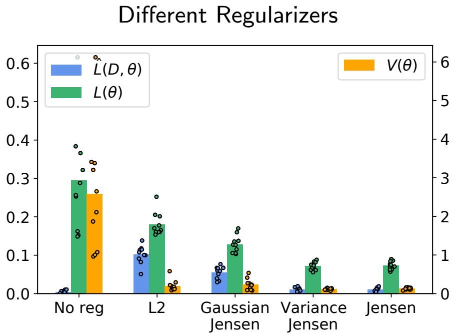
- If the data is normalized and features are conditionally independent, it is **equal to L2-regularization** ,

$$\boldsymbol{\theta}^T \mathsf{Cov}_\nu(y\mathbf{x})\boldsymbol{\theta} = \boldsymbol{\theta}^T kI\boldsymbol{\theta} = k||\boldsymbol{\theta}||^2$$

- **Explains** why L2-regularization improves generalization:
  - Small-norm models tends to have **lower variance**.
  - Lower variance implies **better estimators** $\hat{L}(D, \boldsymbol{\theta})$.
  - Better estimators leads to **less overfitting**.

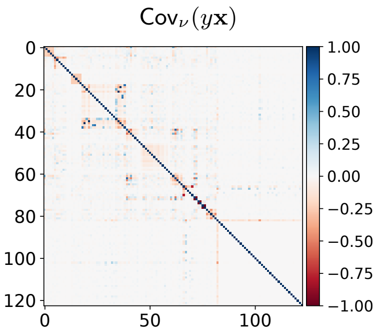
- Also explains the **limitations** of L2-regularization:
  - L2-regularization does not take into account parameter correlations.

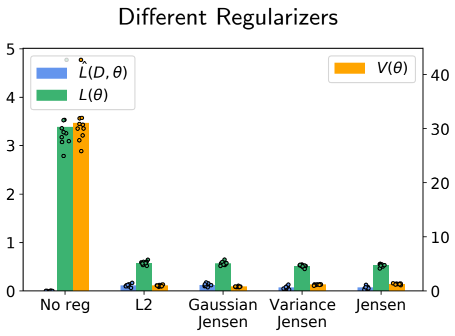




Mushroom Dataset:

- Attributes are highly conditionally (un)correlated.
- $\mathrm{Cov}_\nu(y\mathbf{x})$ very different from a identity matrix.
- L2 performs poorly.





Adult Dataset:

- Attributes are not conditionally correlated.
- $\mathrm{Cov}_\nu(y\mathbf{x})$ very similar to identity matrix.
- L2 performs well.

$\pi_J(\boldsymbol{\theta})$ and existing regularization methods.

More connections with existing regularizations

- For linear regression models, $\pi_J(\boldsymbol{\theta})$ is directly related to **g-priors** (Zellner, 1986).
- In general, $\pi_J(\boldsymbol{\theta})$ is directly related to **input gradient-normalization** (Drucker et al., 1992, Varga et al., 2017).
- Working with more connections with other regularization techniques.

Conclusions and Future Works

Conclusions and Future/Ongoing Works

- PAC-Bayesian bounds and the **generalization performance** of Bayesian methods.
  - **Generalization** is a key property in machine learning.
  - We are **not interested** in finding the best parameters (Bayesian's main goal).
- PAC-Bayesian bounds allow to **identify and correct weaknesses** of Bayesian methods.
  - When learning under model misspecification, Bayesian posterior is not optimal.
  - We can get better performance for the same price.
- PAC-Bayesian bounds allow to better understand **Bayesian priors**.
  - Open problem in Bayesian statistics.
  - We can explain the role of regularizing and informative priors.
  - Explain why (some) regularization methods work.

Conclusions and Future/Ongoing Works

- PAC-Bayesian bounds and the **generalization performance** of Bayesian methods.
  - **Generalization** is a key property in machine learning.
  - We are **not interested** in finding the best parameters (Bayesian's main goal).

- PAC-Bayesian bounds allow to **identify and correct weaknesses** of Bayesian methods.
  - When learning under model misspecification, Bayesian posterior is not optimal.
  - We can get better performance for the same price.

- PAC-Bayesian bounds allow to better understand **Bayesian priors**.
  - Open problem in Bayesian statistics.
  - We can explain the role of regularizing and informative priors.
  - Explain why (some) regularization methods work.

- PAC-Bayesian bounds allow to **better understand ensembles**.
  Ortega et al. Diversity and Generalization in Neural Network Ensembles. AISTATS 2022.

Conclusions and Future/Ongoing Works

- PAC-Bayesian bounds and the **generalization performance** of Bayesian methods.
  - **Generalization** is a key property in machine learning.
  - We are **not interested** in finding the best parameters (Bayesian's main goal).

- PAC-Bayesian bounds allow to **identify and correct weaknesses** of Bayesian methods.
  - When learning under model misspecification, Bayesian posterior is not optimal.
  - We can get better performance for the same price.

- PAC-Bayesian bounds allow to better understand **Bayesian priors**.
  - Open problem in Bayesian statistics.
  - We can explain the role of regularizing and informative priors.
  - Explain why (some) regularization methods work.

- PAC-Bayesian bounds allow to **better understand ensembles**.
  Ortega et al. Diversity and Generalization in Neural Network Ensembles. AISTATS 2022.

- **Future/Ongoing works**:
  - Explain the **Cold Posterior Effect** (Wenzel et al., 2020).

- PAC-Bayesian bounds and the **generalization performance** of Bayesian methods.
  - **Generalization** is a key property in machine learning.
  - We are **not interested** in finding the best parameters (Bayesian's main goal).

- PAC-Bayesian bounds allow to **identify and correct weaknesses** of Bayesian methods.
  - When learning under model misspecification, Bayesian posterior is not optimal.
  - We can get better performance for the same price.

- PAC-Bayesian bounds allow to better understand **Bayesian priors**.
  - Open problem in Bayesian statistics.
  - We can explain the role of regularizing and informative priors.
  - Explain why (some) regularization methods work.

- PAC-Bayesian bounds allow to **better understand ensembles**.
  Ortega et al. Diversity and Generalization in Neural Network Ensembles. AISTATS 2022.

- **Future/Ongoing works**:
  - Explain the **Cold Posterior Effect** (Wenzel et al., 2020).