# Bayesian Priors and Generalization in Probabilistic Machine Learning

Andrés R. Masegosa
(joint work with Luis A. Ortega (UPM, Spain))

*Aalborg University*
*Copenhagen Campus*
*Denmark*

Danish Data Science 2022

# Bayesian machine learning

## Bayesian machine learning

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

## Probabilistic Model

- **Conditional generative models:** $p(y|\mathbf{x}, \boldsymbol{\theta})$.
- **Generative models:** $p(\mathbf{x}|\boldsymbol{\theta})$.

# The Bayesian approach to Machine Learning

## Probabilistic Model

- **Conditional generative models:** $p(y|\mathbf{x}, \boldsymbol{\theta})$.
- **Generative models:** $p(\mathbf{x}|\boldsymbol{\theta})$.

## Bayesian Posterior

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

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

# The Bayesian approach to Machine Learning

## Probabilistic Model

- **Conditional generative models:** $p(y|\mathbf{x}, \boldsymbol{\theta})$.
- **Generative models:** $p(\mathbf{x}|\boldsymbol{\theta})$.

## Bayesian Posterior

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

- 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 Priors in Bayesian Machine Learning (work in progress)

## 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

## 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

- **Regularizing Priors** (e.g., zero centered Gaussian distributions)
  - Promote small norm parameter that reduce overfitting.
  - Overwhelming empirical evidence.
  - Connections with L2, L1, etc. through MAP learning.
- What are Regularizing Priors?
  - Reference priors, (Weakly) Informative priors or **something different**.
- How a Bayesian prior should look like to guarantee **generalization performance**?

## Main Conclusion (Take Away Message)

- Regularizing Priors introduce a biased against **high-variance models**.

## Main Conclusion (Take Away Message)

- Regularizing Priors introduce a biased against **high-variance models**.
- For a fixed $\boldsymbol{\theta}$,
  - We treat $D$ as a random variable, $D \sim \nu(y, \mathbf{x})$ (and $D_{\text{Test}} \sim \nu(y, \mathbf{x})$).
  - **Training Loss**: $\hat{L}(D, \boldsymbol{\theta}) = -\frac{1}{n} \ln p(D|\boldsymbol{\theta})$.
  - **Expected Loss**: $L(\boldsymbol{\theta}) = \mathbb{E}_D[-\frac{1}{n} \ln p(D|\boldsymbol{\theta})] = \mathbb{E}_\nu[-\ln p(y \mid \mathbf{x}, \boldsymbol{\theta})]$.
  - **Variance of the training Loss**: $\mathbb{V}(\boldsymbol{\theta}) = \mathbb{V}_\nu(\frac{1}{n} \ln p(D|\boldsymbol{\theta}))$.

## Main Conclusion (Take Away Message)

- Regularizing Priors introduce a biased against **high-variance models**.
- For a fixed $\boldsymbol{\theta}$,
  - We treat $D$ as a random variable, $D \sim \nu(y, \mathbf{x})$ (and $D_{\text{Test}} \sim \nu(y, \mathbf{x})$).
  - **Training Loss**: $\hat{L}(D, \boldsymbol{\theta}) = -\frac{1}{n} \ln p(D|\boldsymbol{\theta})$.
  - **Expected Loss**: $L(\boldsymbol{\theta}) = \mathbb{E}_D[-\frac{1}{n} \ln p(D|\boldsymbol{\theta})] = \mathbb{E}_\nu[-\ln p(y \mid \mathbf{x}, \boldsymbol{\theta})]$.
  - **Variance of the training Loss**: $\mathbb{V}(\boldsymbol{\theta}) = \mathbb{V}_\nu(\frac{1}{n} \ln p(D|\boldsymbol{\theta}))$.
- If $\mathbb{V}(\boldsymbol{\theta})$ is **large**, then $\boldsymbol{\theta}$ is an **unreliable model**.
  - Training loss can be small and Expected loss can be large.

## Main Conclusion (Take Away Message)

- Regularizing Priors introduce a biased against **high-variance models**.
- For a fixed $\boldsymbol{\theta}$,
  - We treat $D$ as a random variable, $D \sim \nu(y, \mathbf{x})$ (and $D_{\text{Test}} \sim \nu(y, \mathbf{x})$).
  - **Training Loss**: $\hat{L}(D, \boldsymbol{\theta}) = -\frac{1}{n} \ln p(D|\boldsymbol{\theta})$.
  - **Expected Loss**: $L(\boldsymbol{\theta}) = \mathbb{E}_D[-\frac{1}{n} \ln p(D|\boldsymbol{\theta})] = \mathbb{E}_\nu[-\ln p(y \mid \mathbf{x}, \boldsymbol{\theta})]$.
  - **Variance of the training Loss**: $\mathbb{V}(\boldsymbol{\theta}) = \mathbb{V}_\nu(\frac{1}{n} \ln p(D|\boldsymbol{\theta}))$.
- If $\mathbb{V}(\boldsymbol{\theta})$ is **large**, then $\boldsymbol{\theta}$ is an **unreliable model**.
  - Training loss can be small and Expected loss can be large.
- **Generalization requires penalizing high-variance models**.
  - This should be encoded in the prior:

$$\pi_J(\boldsymbol{\theta}) \propto e^{-\mathbb{V}(\theta)}$$

  - **Connected** to Gaussian zero-centered priors and other regularizing priors.

## PAC-Bayes Bounds

- Frequentist analysis of Bayesian methods ($D$ is **treated as a random quantity**).
- Theoretical tool to analyse the relationship between training loss and test loss:

$$\text{Expected Loss} \leq \text{Train Loss} + \text{Complexity}$$

## PAC-Bayes Bounds

- Frequentist analysis of Bayesian methods ($D$ is **treated as a random quantity**).
- Theoretical tool to analyse the relationship between training loss and test loss:

$$\text{Expected Loss} \leq \text{Train Loss} + \text{Complexity}$$

- **Expected Loss** of Bayesian learning:

$$\underbrace{CE(p(\boldsymbol{\theta}|D))}_{\text{Bayesian Expected Loss}} = \mathbb{E}_{\nu(\mathbf{y},\mathbf{x})}[-\ln \underbrace{\mathbb{E}_{p(\theta|D)}[p(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})]}_{\text{Bayesian Model Averaging}}]$$

  - $\nu(\mathbf{y},\mathbf{x})$ is the data-generating distribution and $D \sim \nu(\mathbf{y},\mathbf{x})$.

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

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

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

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

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

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

where $p_\pi^\lambda$ denotes the **generalized Bayesian posterior**,

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

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

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

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

where $p_\pi^\lambda$ 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 **(Alquier et al. 2016, Germain et al. 2016, Masegosa 2020)**

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

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

where $p_\pi^\lambda$ 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:

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

## Upper Bounds

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

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

## Upper Bounds

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

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

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

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

## Upper Bounds

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

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

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

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

- According to these bounds, **small predictive loss is attained if**:
  - $-L\hat{M}_\lambda(\pi, D)$ **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:

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

## 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:

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

- Then, we have that

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

- 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 $\lambda > 0$, ,

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

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

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

## Upper Bounds

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

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

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

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

- **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

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

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

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

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 $\lambda > 0$, ,

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

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

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

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

## Theorem: Optimal Priors

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

$$\pi_1(\boldsymbol{\theta}) \propto \underbrace{\pi_I(\boldsymbol{\theta})}_{\text{Informative Prior}} \underbrace{e^{-nJ_\nu(\theta,\lambda)}}_{\text{Regularizing Prior}}$$

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:

$$\pi_1(\boldsymbol{\theta}) \propto \underbrace{\pi_I(\boldsymbol{\theta})}_{\text{Informative Prior}} \underbrace{e^{-nJ_\nu(\theta,\lambda)}}_{\text{Regularizing Prior}}$$

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

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

## 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 Expected Loss.

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

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

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

- $\pi_J(\boldsymbol{\theta})$ *addresses* overfitting (i.e., a regularizing prior).
  - It is a **functional prior**.
  - The prior depends on $p(y \mid \mathbf{x}, \boldsymbol{\theta})$ and the $\nu(y, \mathbf{x})$.

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

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

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

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

$$\theta_{\text{MAP}} = \arg\max_{\theta} p^{\lambda}_{\pi_J}(\boldsymbol{\theta}|D)$$

$$= \arg\min_{\theta} \hat{L}(\boldsymbol{\theta}, D) - \underbrace{\frac{\ln \pi_J(\boldsymbol{\theta})}{\lambda n}}_{\text{log Prior}}$$

$$= \arg\min_{\theta} \hat{L}(\boldsymbol{\theta}, D) + \underbrace{\frac{J_\nu(\boldsymbol{\theta}, \lambda)}{\lambda}}_{\text{Regularizer}}$$

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

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

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

- 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 $\boldsymbol{\theta}$:

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

## $\pi_J(\boldsymbol{\theta})$ 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 $\boldsymbol{\theta}$:

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

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

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

## $\pi_J(\boldsymbol{\theta})$ 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 $\boldsymbol{\theta}$:

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

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

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

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

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

## $\pi_J(\boldsymbol{\theta})$ 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 $\boldsymbol{\theta}$:

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

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

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

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

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

- **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 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 $\boldsymbol{\theta}$:

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

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

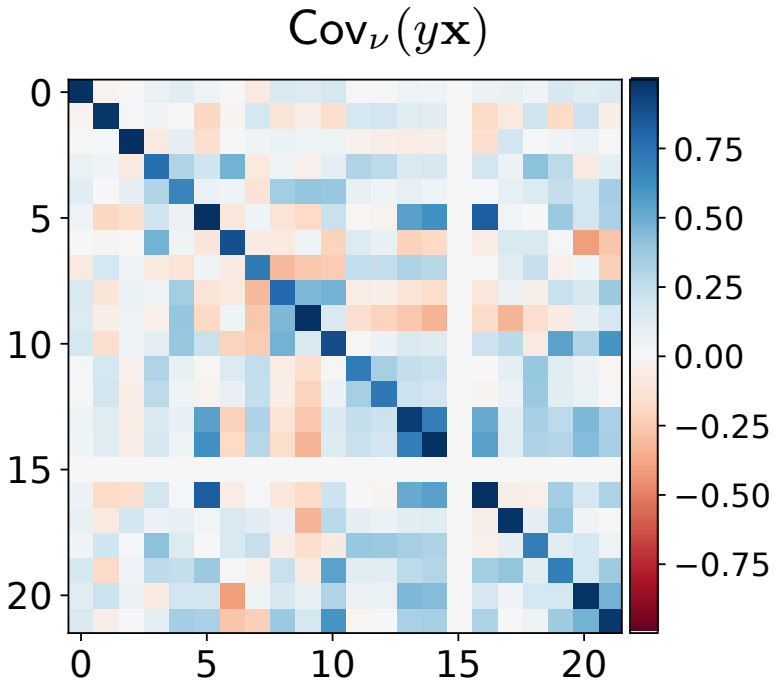
$$\pi_J(\boldsymbol{\theta}) \propto e^{-n0.25\lambda^2\theta^T \text{Cov}_\nu(y\mathbf{x})\theta}$$

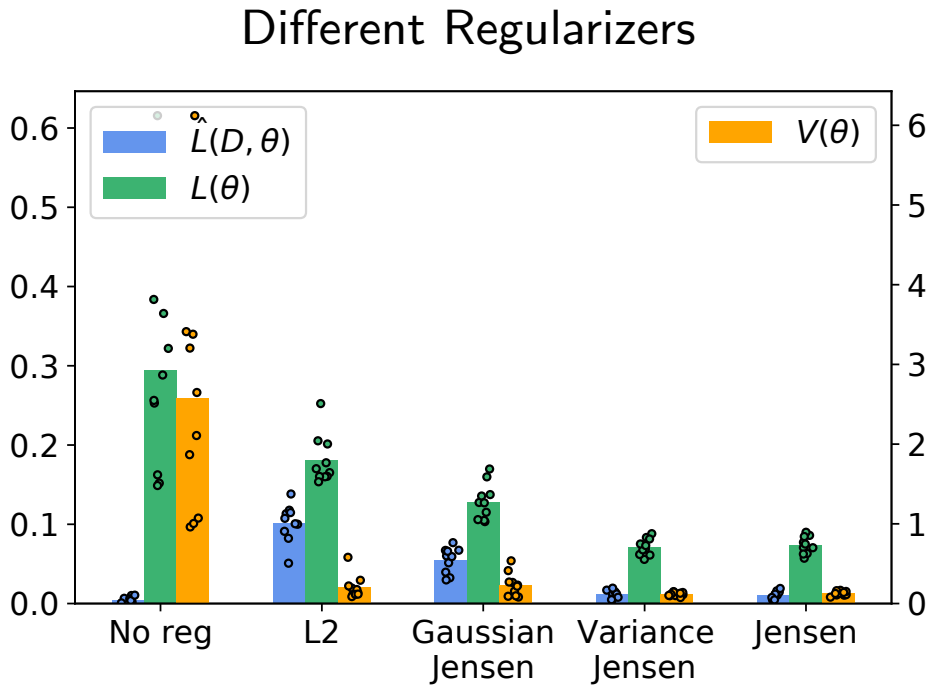
- If the data is normalized and features are conditionally independent, it is **equal to L2-regularization** ,

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

- **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**.

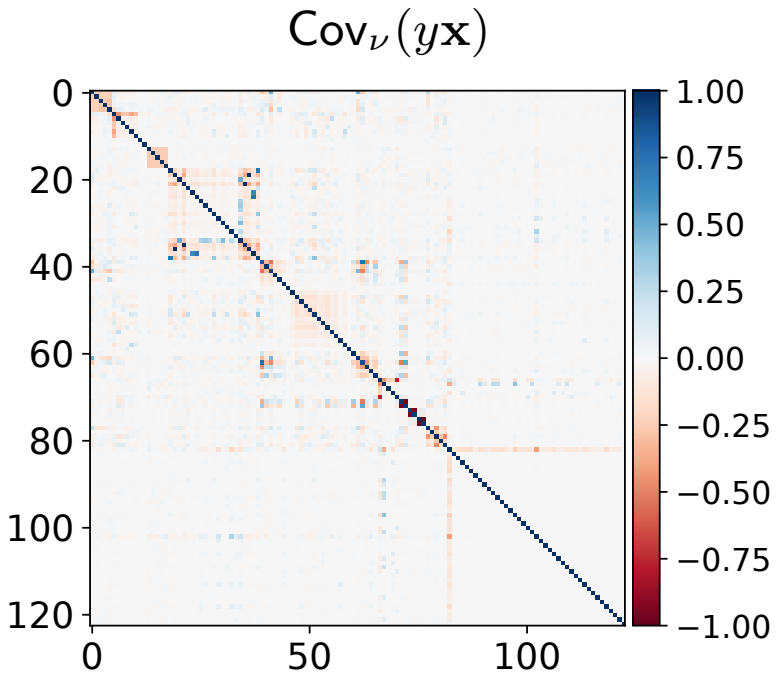
- Also explains the **limitations** of L2-regularization:
  - L2-regularization does not take into account parameter correlations.

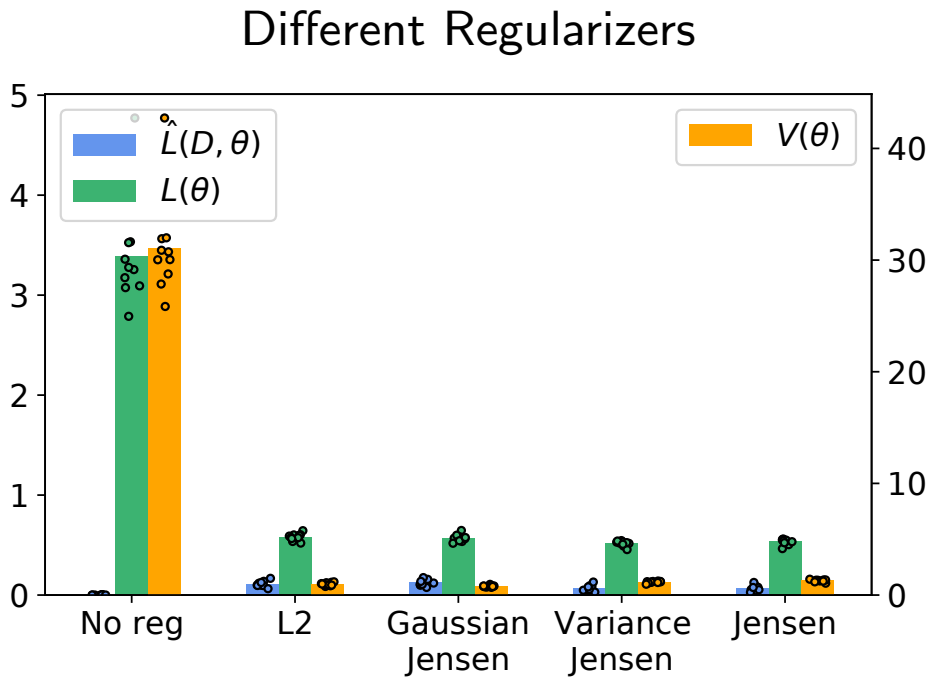



**Mushroom Dataset:**

- 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).
- $\pi_J(\boldsymbol{\theta})$ 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.

- 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 **identify and correct weaknesses** of Bayesian methods.
  - When learning under model misspecification, Bayesian posterior is not optimal (Masegosa, 2020).
  - We can get better performance for the same price.

# Conclusions and Future/Ongoing Works

- PAC-Bayesian bounds and the **generalization performance** of Bayesian methods.
  - **Generalization** is a key property in machine learning.

- 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 **identify and correct weaknesses** of Bayesian methods.
  - When learning under model misspecification, Bayesian posterior is not optimal (Masegosa, 2020).
  - We can get better performance for the same price.

- 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.

- 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 **identify and correct weaknesses** of Bayesian methods.
  - When learning under model misspecification, Bayesian posterior is not optimal (Masegosa, 2020).
  - We can get better performance for the same price.

- 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).

# Conclusions and Future/Ongoing Works

- PAC-Bayesian bounds and the **generalization performance** of Bayesian methods.
  - **Generalization** is a key property in machine learning.

- 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 **identify and correct weaknesses** of Bayesian methods.
  - When learning under model misspecification, Bayesian posterior is not optimal (Masegosa, 2020).
  - We can get better performance for the same price.

- 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).