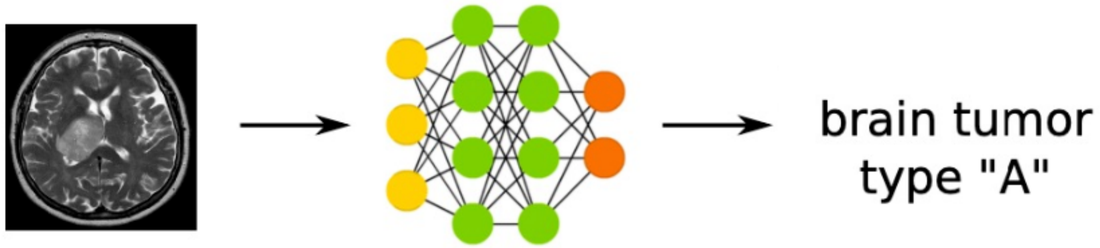
A Generalisation Analysis of Bayesian Machine Learning

Andres Masegosa
Aalborg University

Uncertainty in Machine Learning



- How confident is a ML model when making an specific prediction?
  - Quite relevant in many domains.

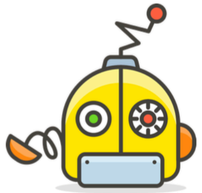
Image from: Yingzhen Li. An Introduction to Bayesian Neural Networks. ProbAI 2023

Uncertainty in Machine Learning

Desiderata

Here's this patient's health record:
...
Could you summarise it for me?

Here's a summary of the patient's health record you requested:
[point 1] with x% confidence
...



Here's the conditions of this construction site:
...
Could you tell me what the potential safety issues are?

Here's potential safety issues that need to be look after:
[point 1] with x% confidence
...

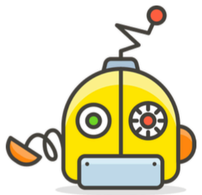
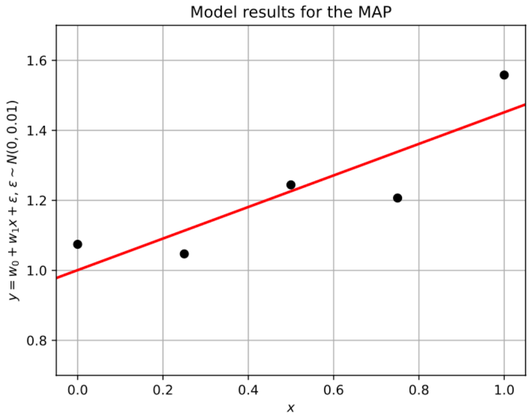


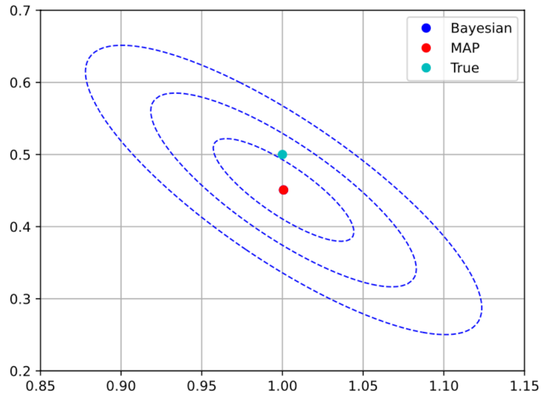
Image from: Yingzhen Li. An Introduction to Bayesian Neural Networks. ProbAI 2023

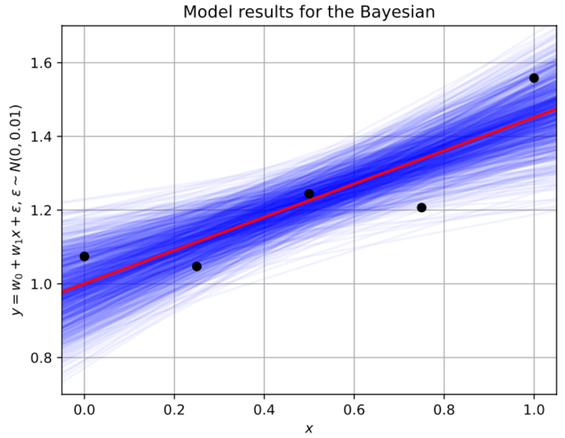
Bayesian in Machine Learning

- Bayesian methods are **widely used** in machine learning for uncertainty quantification.
- 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 Linear Regression







$$\theta_{MAP} = \arg\max_{\theta} p(D \mid \theta)$$

$$\theta_{MAP} = \arg\max_{\theta} p(\theta \mid D)$$

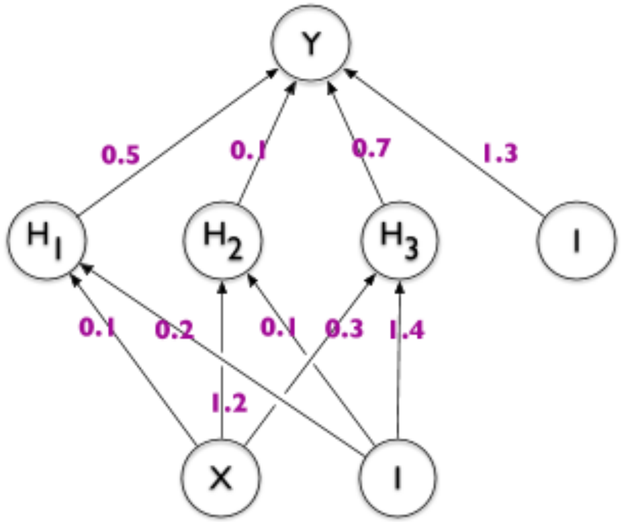
$$\theta_i \sim p(\theta \mid D)$$

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

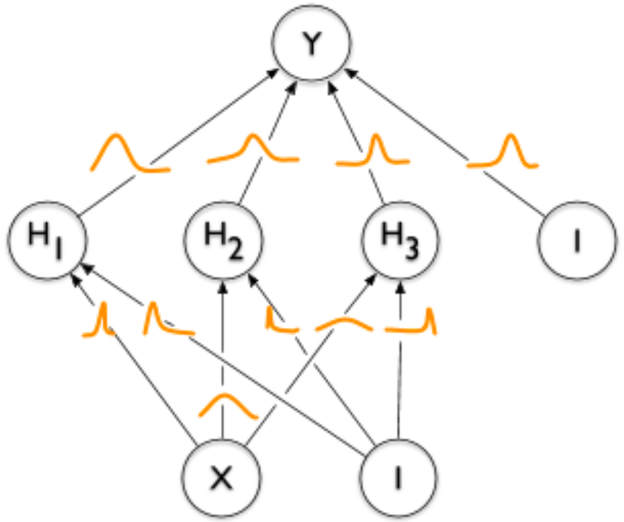
Image from: Langseth et al. Variational Inference and Optimization. ProbAI 2023

Bayesian Neural Networks

θ_{MAP}

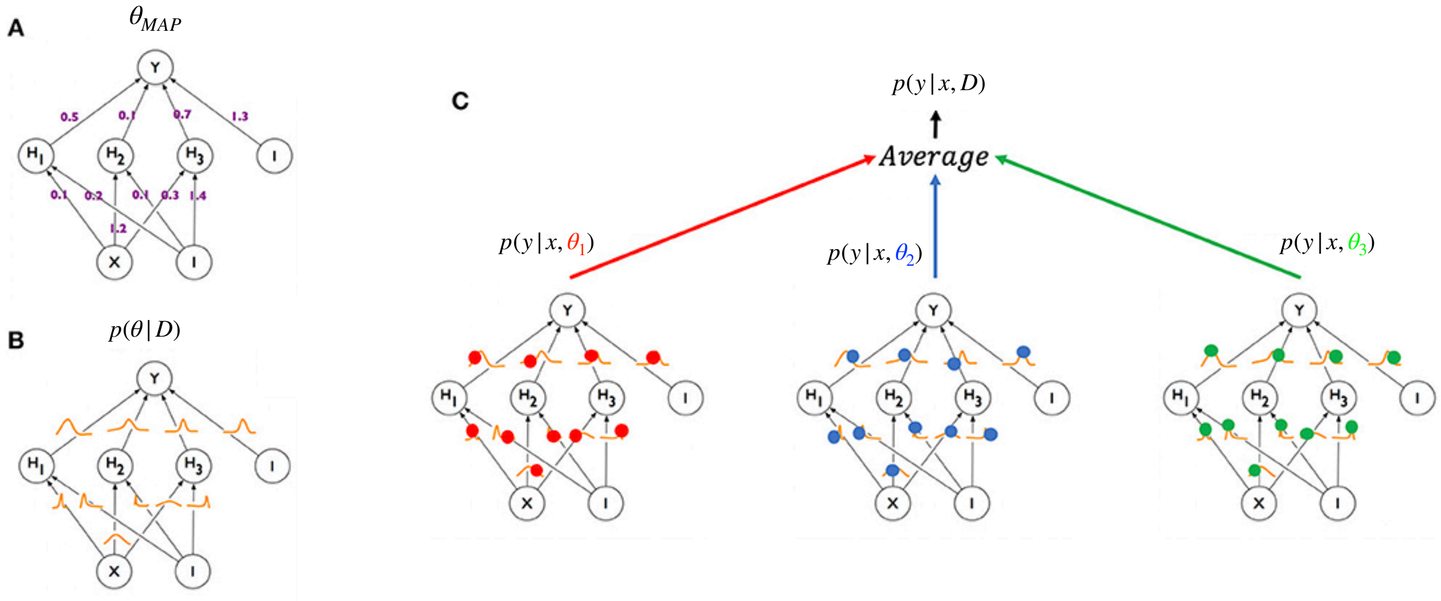


$p(\theta \mid D)$



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

Bayesian Neural Networks



$$\theta_i \sim p(\theta \mid D)$$

Main Focus of Bayesian Machine Learning

- The main problem is how to compute the **Bayesian posterior**

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

  - Intractable for models with a large number of parameters.
  - There exist many approximate inference schemes.

Do Bayesian NNs have better performance?

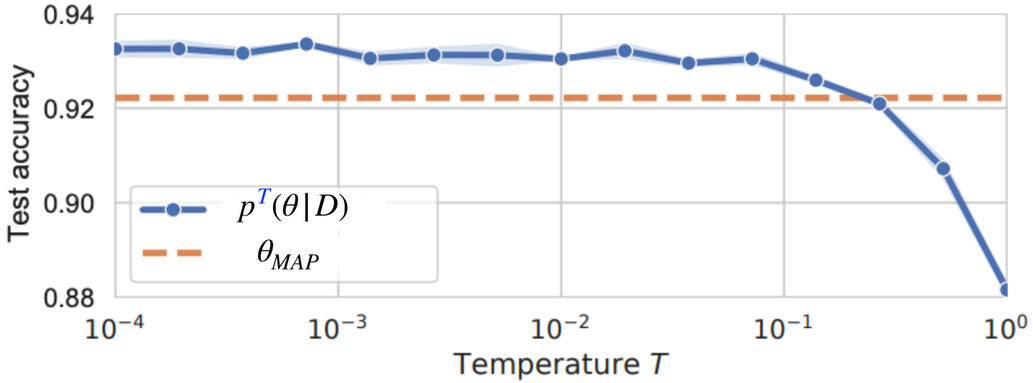
The answer is “not yet”.

Cold Posterior Effect (Wenzel et al., 2020)

- **Tempered Posteriors:**

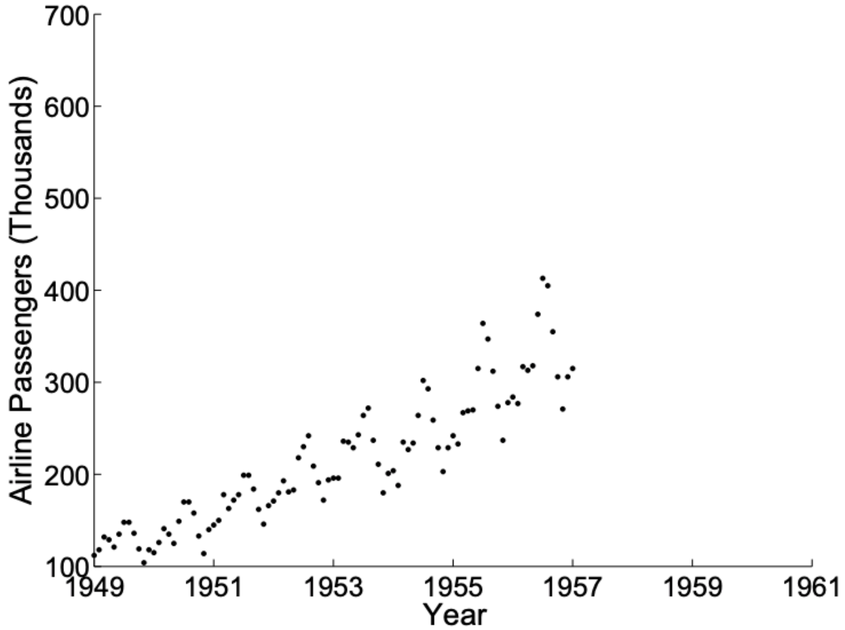
$$p^T(\theta \mid D) \propto (P(D \mid \theta)\pi(\theta))^{1/T}$$

  - T=1 corresponds to the Bayesian posterior.
  - T=1 should be optimal for predictive performance.



- **Why do Cold Posteriors perform better?**
  - Approximate inference methods (Izmailov et al., 2021).
  - Model Misspecification issue (Aitchison, 2021; Kapoor et al., 2022; Fortuin et al., 2022).
  - Underfitting issue (Zhang et al. 2023).

Model Selection



- Which model should we choose?

$$p_1(y \mid x, \theta) = \mathcal{N}(\theta_0 + \theta_1 x, \sigma)$$

$$p_2(y \mid x, \theta) = \mathcal{N}(\sum_{j=0}^{3} \theta_i x, \sigma)$$

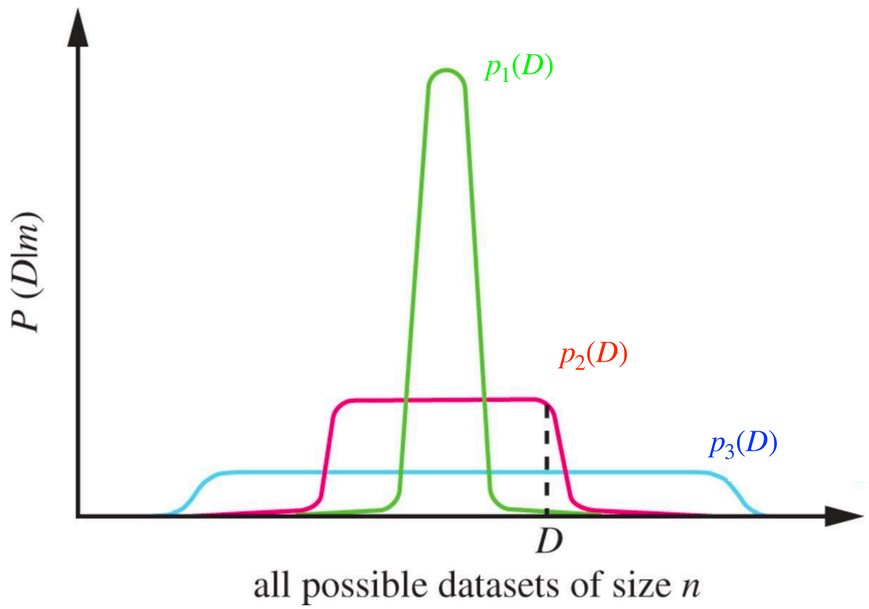
$$p_3(y \mid x, \theta) = \mathcal{N}(\sum_{j=0}^{10^4} \theta_i x, \sigma)$$

Image from: Andrew G. Wilson. Bayesian Deep Learning and a Probabilistic Perspective of Model Construction. ICML 2020.

Bayesian Model Selection (Mackay, 2003)

- **Marginal Likelihood**

$$p_k(D) = \int p_k(D \mid \theta)\pi_k(\theta)d\theta$$

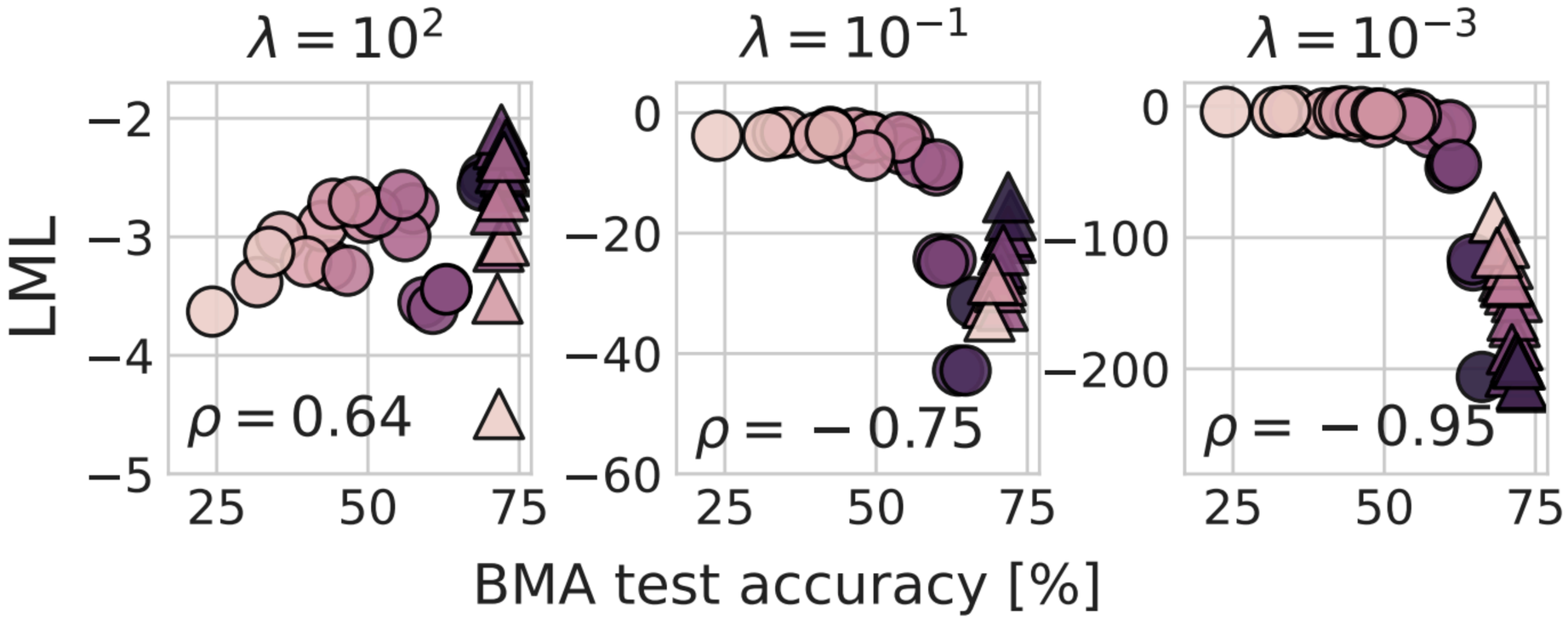


$$p_1(y \mid x, \theta) = \mathcal{N}(\theta_0 + \theta_1 x, \sigma) \qquad p_2(y \mid x, \theta) = \mathcal{N}(\sum_{j=0}^{3} \theta_i x, \sigma) \qquad p_3(y \mid x, \theta) = \mathcal{N}(\sum_{j=0}^{10^4} \theta_i x, \sigma)$$

Does Marginal Likelihood work for NNs?

The answer is “not yet”.

Marginal Likelihood in NNs (Lotfi et al. 2023)



- Marginal Likelihood is poorly correlated with generalisation

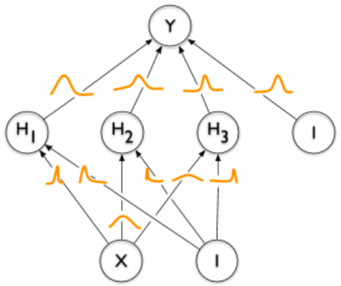
A Generalisation Analysis of Bayesian Machine Learning

PAC-Bayesian Analysis

PAC-Bayesian Bounds

(Alquier et al. 2016, Germain et al. 2016, Masegosa 2020)

$$\theta \sim \rho(\theta)$$



- 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**, $\hat{L}(\boldsymbol{\theta}, D) = -\frac{1}{n}\ln p(D|\boldsymbol{\theta})$
  - $B(\rho)$ is the **Bayes loss** of the posterior ρ, $B(\rho) = \mathbb{E}_{\nu(\mathbf{y},\mathbf{x})}[-\ln \mathbb{E}_{\rho}[p(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})]]$

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

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

  - $\phi_{\nu}^{\pi}(\lambda)$ is a **log moment generating function**.

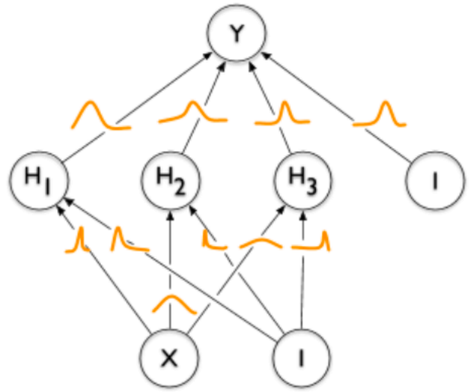
PAC-Bayesian Bounds

(Alquier et al. 2016, Germain et al. 2016, Masegosa 2020)

$$\theta \sim \rho(\theta)$$

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

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



- Which is the optimal density $\rho(\boldsymbol{\theta})$?,

$$\rho^\star(\boldsymbol{\theta}|D) \quad = \quad \arg\min_\rho \underbrace{\mathbb{E}_\rho[\hat{L}(\boldsymbol{\theta}, D)] + \frac{KL(\rho, \pi) + \phi_\nu^\pi(\lambda) + \ln \frac{1}{\delta}}{\lambda n}}_{\text{PAC-Bayes bound}}$$

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

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

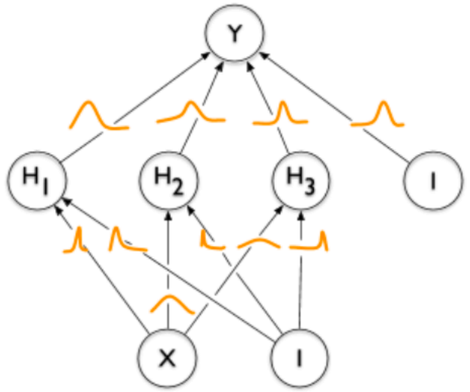
PAC-Bayesian Bounds

(Alquier et al. 2016, Germain et al. 2016, Masegosa 2020)

$$\theta \sim \rho(\theta)$$

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

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



- Which is the optimal density $\rho(\boldsymbol{\theta})$?,

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

PAC-Bayesian Bounds

(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,

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

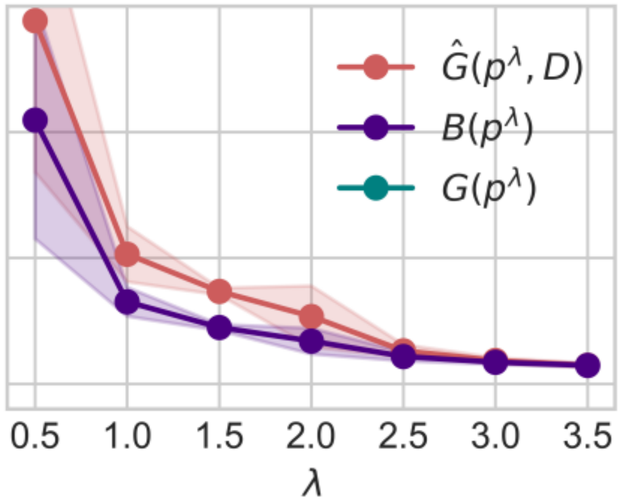
- Which is the optimal density $\rho(\boldsymbol{\theta})$?,

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

- $p^{\lambda}(\boldsymbol{\theta}|D)$ is the **Tempered Bayesian posterior**,

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

Cold Posterior Effect



(Zhang et al. 2023)

Marginal Likelihood and Generalization

- The Log Marginal Likelihood decomposes as:

$$-\frac{1}{n}\ln p(D) = \mathbb{E}_p[\hat{L}(D,\boldsymbol{\theta})] + \frac{KL(p,\pi)}{n}$$

- For the Bayesian posterior we have $p(\theta|D)$,

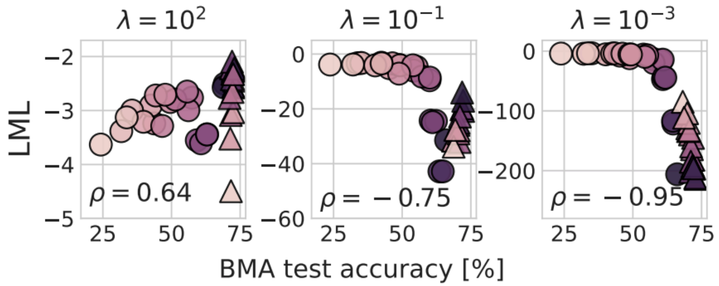
$$\underbrace{B(p)}_{\text{Bayes Loss}} \overbrace{\leq}^{\text{Jensen Inequality}} \underbrace{\mathbb{E}_p[L(\boldsymbol{\theta})]}_{\text{Gibbs Loss}} \overbrace{\lesssim}^{\text{w.p. } (1-\delta)} \underbrace{\mathbb{E}_p[\hat{L}(\boldsymbol{\theta},D)] + \phi^\star\big(\frac{KL(p,\pi)+\ln\frac{|\Lambda|}{\delta}}{n}\big)}_{\text{PAC-Bayes bound}}$$

  - $\phi^\star(a)$ is an approximation of the the inverse Legendre transform of ϕ^π_ν.

- The Log Marginal would be a good proxy of generalization when $\phi^\star(a) = a$.

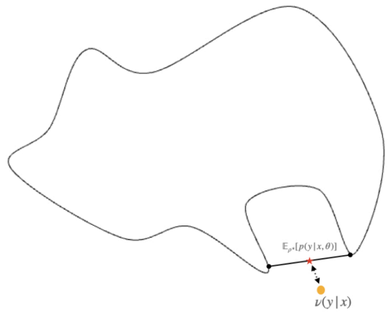
$$\underbrace{B(p)}_{\text{Bayes Loss}} \overbrace{\leq}^{\text{Jensen Inequality}} \underbrace{\mathbb{E}_p[L(\boldsymbol{\theta})]}_{\text{Gibbs Loss}} \overbrace{\lesssim}^{\text{w.p. } (1-\delta)} \underbrace{\mathbb{E}_p[\hat{L}(\boldsymbol{\theta},D)] + \frac{KL(p,\pi)+\ln\frac{|\Lambda|}{\delta}}{n}}_{\text{PAC-Bayes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

LML vs Generalisation



Model Misspecification (Masegosa, 2020)

$$\arg\min_{\rho} \underbrace{B(\rho)}_{\text{Bayes Loss}} \neq \arg\min_{\rho} \underbrace{\mathbb{E}_{\rho}[L(\boldsymbol{\theta})]}_{\text{Gibbs Loss}}$$



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

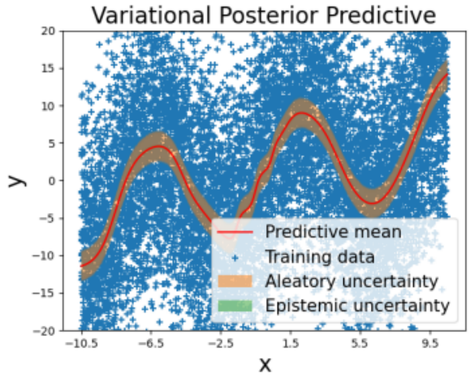
$$\underbrace{B(\rho)}_{\text{Bayes Loss}} \overset{\text{Jensen Inequality}}{\leq} \underbrace{\mathbb{E}_{\rho}[L(\boldsymbol{\theta})]}_{\text{Gibbs Loss}} \overset{\text{w.p. } (1-\delta)}{\lesssim} \underbrace{\mathbb{E}_{\rho}[\hat{L}(\boldsymbol{\theta}, D)] + \frac{KL(\rho, \pi) + \phi_{\nu}(n,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

The Bayesian Posterior is not optimal under model misspecification

Bayesian Learning under Model Misspecification

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

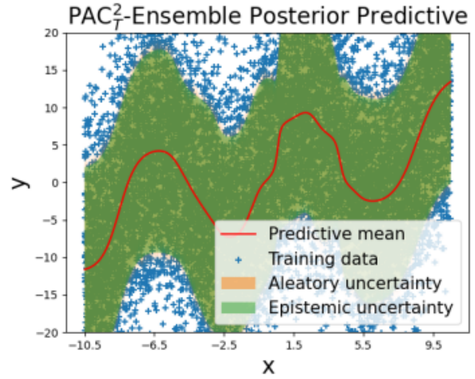
$$\underbrace{B(\rho)}_{\text{Bayes Loss}} \leq \underbrace{\mathbb{E}_{\theta^{(m)}\sim\rho}[L_P(\boldsymbol{\theta}^{(m)})]}_{\text{Extended Gibbs Loss}} \overset{\text{w.p. } (1-\delta)}{\lesssim} \underbrace{\mathbb{E}_{\theta^{(m)}\sim\rho}[\overbrace{\hat{L}_P(\boldsymbol{\theta}^{(m)}, D)}^{\text{Empirical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



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

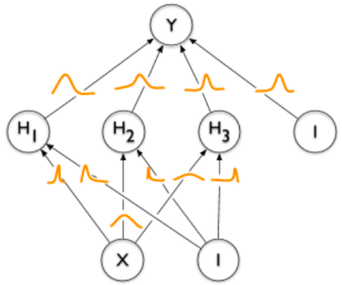
$$\underbrace{B(\rho)}_{\text{Bayes Loss}} \leq \underbrace{\mathbb{E}_{\theta^{(m)}\sim\rho}[L_P(\boldsymbol{\theta}^{(m)})]}_{\text{Extended Gibbs Loss}} \leq \underbrace{\mathbb{E}_{\rho}[L(\boldsymbol{\theta})]}_{\text{Gibbs Loss}}$$

  - (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).



PAC-Bayesian Bounds

$$\theta \sim p(\theta \,|\, D)$$



- Many open questions in Bayesian Machine learning:
  - What is the Cold Posterior Effect?
  - When does the marginal likelihood correlate with generalization?
  - How to design better Bayesian priors?
  - How to build better learning algorithms?

- PAC-Bayesian bounds provide a generalization perspective of Bayesian methods.
  - Potential to achieve a better understanding of BML.
  - Potential to design better learning algorithms.

- Open Issues with PAC-Bayes:
  - The log-loss used in Bayesian learning is unbounded.
  - Are current PAC-Bayes bounds tight?
  - Do we need distribution-dependent PAC-Bayes bounds?
  - Should we move to information theoretic bounds?
  - Derive a *killer* learning algorithm.

Thanks for you attention!! :D