Bayesian Model Averaging is not Model Combination: A PAC-Bayesian Analysis of Deep Ensembles

Andrés R. Masegosa

University of Aalborg (Copenhagen Campus)
Denmark

Masegosa, A. R. Learning under model misspecification: Applications to variational and ensemble methods. NeurIPS 2020.
Ortega, L. A., Cabañas R., Masegosa, A. R. Diversity and Generalization in Neural Network Ensembles. arxiv/2110.13786. 2021

Bayesian Model Averaging vs Model Combination

Bayesian model averaging is not model combination

Thomas P. Minka

December 13, 2002

In a recent paper, Domingos (2000) compares Bayesian model averaging (BMA) to other model combination methods on some benchmark data sets, is surprised that BMA performs worst, and suggests that BMA may be flawed. These results are actually *not* surprising, especially in light of an earlier paper by Domingos (1997) where it was shown that model combination works by enriching the space of hypotheses, not by approximating a Bayesian model average. And the only flaw with BMA is the belief that it is an algorithm for model combination, when it is not.

Model Combination

- Model Combination works by **enriching the model space**.
- BMA represents the **inability to distinguish the best single model when using limited data**.

Deep Ensembles (Lakshminarayanan et al., 2017, Wenzel et al., 2020)

- **Deep Ensembles** provides SOTA performance in terms of:
  - Uncertainty Estimation.
  - Robustness Against Distributional Shifts.

Deep Ensembles (Lakshminarayanan et al., 2017, Wenzel et al., 2020)

- **Deep Ensembles** provides SOTA performance in terms of:
  - Uncertainty Estimation.
  - Robustness Against Distributional Shifts.

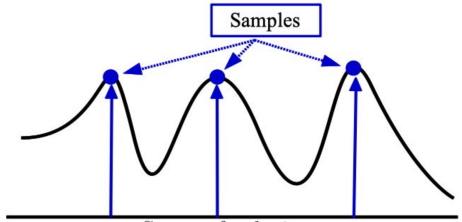
Open Question: Why do deep ensembles work so well?

Deep Ensembles as Approximate Bayesian Inference

Andrew Gordon Wilson, Pavel Izmailov
October 21, 2021

Science is about updating our beliefs based on new evidence. It sounds simple and obvious, but this task is often extraordinarily difficult in practice. In reality, it's easy to get invested in a particular narrative. How often have you encountered reviewers who refuse to update their assessments, even when their criticisms are fully addressed, or shown to be misguided? In those situations, how often does it feel like the goal posts are being moved around, just to maintain the same underlying narrative, even if the arguments are changing?

Deep ensembles are a popular approach that works by retraining the same neural network multiple times, and averaging the resulting models. When deep ensembles were introduced broadly to the machine learning community, they were often presented as a non-Bayesian competitor to various approximate Bayesian inference procedures, particularly mean-field variational inference. "Non-Bayesian" deep ensembles typically outperformed these "Bayesian" methods, both in accuracy and calibration. The common takeaway? We should forget about Bayesian methods in deep learning, for they don't work well in practice. "Why not just have a workshop about deep ensembles?", was a common question at the annual Bayesian deep learning workshop, at NeurIPS 2019.



$$\underbrace{\mathbb{E}_{p(\theta|D)}[p(y \mid \mathbf{x}, \boldsymbol{\theta})]}_{\text{Bayesian Predictive posterior}} \approx \underbrace{\frac{1}{K}\sum_{i} \overbrace{p(y \mid \mathbf{x}, \boldsymbol{\theta}_i)}^{\text{Model Prediction}}}_{\text{Finite Sample Approximation}}$$

Deep Ensembles (Lakshminarayanan et al., 2017, Wenzel et al., 2020)

- **Deep Ensembles** provides SOTA performance in terms of:
  - Uncertainty Estimation.
  - Robustness Against Distributional Shifts.

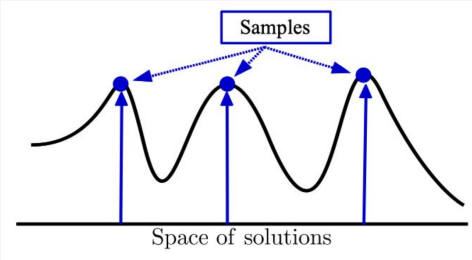
Open Question: Why do deep ensembles work so well?

Deep Ensembles as Approximate Bayesian Inference

Andrew Gordon Wilson, Pavel Izmailov
October 21, 2021

Science is about updating our beliefs based on new evidence. It sounds simple and obvious, but this task is often extraordinarily difficult in practice. In reality, it's easy to get invested in a particular narrative. How often have you encountered reviewers who refuse to update their assessments, even when their criticisms are fully addressed, or shown to be misguided? In those situations, how often does it feel like the goal posts are being moved around, just to maintain the same underlying narrative, even if the arguments are changing?

Deep ensembles are a popular approach that works by retraining the same neural network multiple times, and averaging the resulting models. When deep ensembles were introduced broadly to the machine learning community, they were often presented as a non-Bayesian competitor to various approximate Bayesian inference procedures, particularly mean-field variational inference. "Non-Bayesian" deep ensembles typically outperformed these "Bayesian" methods, both in accuracy and calibration. The common takeaway? We should forget about Bayesian methods in deep learning, for they don't work well in practice. "Why not just have a workshop about deep ensembles?", was a common question at the annual Bayesian deep learning workshop, at NeurIPS 2019.



$$\underbrace{\mathbb{E}_{p(\theta|D)}[p(y \mid \mathbf{x}, \boldsymbol{\theta})]}_{\text{Bayesian Predictive posterior}} \approx \underbrace{\frac{1}{K}\sum_{i} \overbrace{p(y \mid \mathbf{x}, \boldsymbol{\theta}_i)}^{\text{Model Prediction}}}_{\text{Finite Sample Approximation}}$$

- Benefits of deep ensembles are **due to their Bayesian nature**.
- Deep Ensembles do not perform **model combination** (Minka, 2002).

- Let us assume we have:
  - **Model class**: $\{p(y|\mathbf{x}, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$. (e.g. neural network fixed architecture).
  - **Training Data Sample** D.

- **Bayesian posterior**:

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

- Let us assume we have:
  - **Model class**: $\{p(y|\mathbf{x}, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$. (e.g. neural network fixed architecture).
  - **Training Data Sample** D.

- **Bayesian posterior**:

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

- **Bayesian model averaging (BMA)**:

$$\underbrace{\mathbb{E}_{p(\theta|D)}[p(y \mid \mathbf{x}, \boldsymbol{\theta})]}_{\text{Bayesian Predictive posterior}} = \int \overbrace{p(y \mid \mathbf{x}, \boldsymbol{\theta})d\boldsymbol{\theta}}^{\text{Model Prediction}} \quad \underbrace{p(\boldsymbol{\theta}|D)}_{\text{Bayesian posterior}} \quad d\boldsymbol{\theta}$$

- $\rho(\boldsymbol{\theta})$ denotes a **quasi-posterior**:
  - Defines a probability distribution over $\boldsymbol{\Theta}$.
  - May depends on the training data sample D.

Model Combination

- $\rho(\boldsymbol{\theta})$ denotes a **quasi-posterior**:
  - Defines a probability distribution over $\boldsymbol{\Theta}$.
  - May depends on the training data sample D.

- **Model Combination** defined by $\rho(\boldsymbol{\theta})$:

$$\underbrace{\mathbb{E}_{\rho}[p(y \mid \mathbf{x}, \boldsymbol{\theta})]}_{\text{Predictive posterior}} = \int \overbrace{p(y \mid \mathbf{x}, \boldsymbol{\theta})}^{\text{Model Prediction}} \underbrace{\rho(\boldsymbol{\theta})}_{\text{Quasi-posterior}} d\boldsymbol{\theta}$$

- $\rho(\boldsymbol{\theta})$ denotes a **quasi-posterior**:
  - Defines a probability distribution over $\boldsymbol{\Theta}$.
  - May depends on the training data sample D.

- **Model Combination** defined by $\rho(\boldsymbol{\theta})$:

$$\underbrace{\mathbb{E}_{\rho}[p(y \mid \mathbf{x}, \boldsymbol{\theta})]}_{\text{Predictive posterior}} = \int \overbrace{p(y \mid \mathbf{x}, \boldsymbol{\theta})}^{\text{Model Prediction}} \underbrace{\rho(\boldsymbol{\theta})}_{\text{Quasi-posterior}} d\boldsymbol{\theta}$$

- **Considerations**:
  - BMA is special case of model combination (i.e. when $\rho(\boldsymbol{\theta}) = p(\boldsymbol{\theta} \mid D)$).
  - We can choose any distribution ρ over $\boldsymbol{\Theta}$ using information from D.

Generalization Performance of Model Combination (Masegosa 2020)

- We assume there is an **unknown data generating distribution** $\nu(y, \mathbf{x})$.
  - The training/test data are generated from ν (i.e $D \sim \nu(y, \mathbf{x})$).

Generalization Performance of Model Combination (Masegosa 2020)

- We assume there is an **unknown data generating distribution** $\nu(y, \mathbf{x})$.
  - The training/test data are generated from ν (i.e $D \sim \nu(y, \mathbf{x})$).

- **Generalization Error of a ρ-combined model** using cross-entropy loss.

$$\underbrace{L_{ce}(\rho)}_{\substack{\textbf{Gen. Error of} \\ \rho\text{-combined model}}} = \mathbb{E}_{\nu}[-\ln \underbrace{\mathbb{E}_{\rho}[p(y \mid \mathbf{x}, \boldsymbol{\theta})]}_{\substack{\text{Predictive posterior} \\ \text{of } \rho(\boldsymbol{\theta})}}]$$

Generalization Performance of Model Combination (Masegosa 2020)

- We assume there is an **unknown data generating distribution** $\nu(y, \mathbf{x})$.
  - The training/test data are generated from ν (i.e $D \sim \nu(y, \mathbf{x})$).

- **Generalization Error of a ρ-combined model** using cross-entropy loss.

$$\underbrace{L_{ce}(\rho)}_{\substack{\textbf{Gen. Error of} \\ \rho\text{-combined model}}} = \mathbb{E}_{\nu}[-\ln \underbrace{\mathbb{E}_{\rho}[p(y \mid \mathbf{x}, \boldsymbol{\theta})]}_{\substack{\text{Predictive posterior} \\ \text{of } \rho(\boldsymbol{\theta})}}]$$

- Model combination works with **an enriched model class** parametrized by ρ:

$$\underbrace{\{p(y|\mathbf{x}, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}}_{\text{Model class parametrized by } \theta} \subseteq \underbrace{\{\mathbb{E}_{\rho}[p(y \mid \mathbf{x}, \boldsymbol{\theta})] : \rho \text{ is a quasi-posterior over } \boldsymbol{\Theta}\}}_{\text{Model class parametrized by } \rho}$$

Generalization Performance of Model Combination (Masegosa 2020)

- We assume there is an **unknown data generating distribution** $\nu(y, \mathbf{x})$.
  - The training/test data are generated from ν (i.e $D \sim \nu(y, \mathbf{x})$).

- **Generalization Error of a ρ-combined model** using cross-entropy loss.

$$\underbrace{L_{ce}(\rho)}_{\substack{\textbf{Gen. Error of} \\ \rho\text{-combined model}}} = \mathbb{E}_{\nu}[-\ln \underbrace{\mathbb{E}_{\rho}[p(y \mid \mathbf{x}, \boldsymbol{\theta})]}_{\substack{\text{Predictive posterior} \\ \text{of } \rho(\boldsymbol{\theta})}}]$$

- Model combination works with **an enriched model class** parametrized by ρ:

$$\underbrace{\{p(y|\mathbf{x}, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}}_{\text{Model class parametrized by } \theta} \subseteq \underbrace{\{\mathbb{E}_{\rho}[p(y \mid \mathbf{x}, \boldsymbol{\theta})] : \rho \text{ is a quasi-posterior over } \boldsymbol{\Theta}\}}_{\text{Model class parametrized by } \rho}$$

- In consequence

$$\min_{\rho} \underbrace{L_{ce}(\rho)}_{\substack{\textbf{Gen. Error of} \\ \rho\text{-combined model}}} \leq \min_{\theta} \underbrace{L_{ce}(\boldsymbol{\theta})}_{\substack{\textbf{Gen. Error of} \\ \text{the model } \theta}}$$

PAC-Bayesian upper-bounds (Alquier et al. 2016, Germain et al. 2016, Masegosa 2020)

$$\underbrace{L_{ce}(\rho)}_{\text{Gen. Error of } \rho\text{-combined model}} \overbrace{\leq}^{\text{Jensen Inequality}} \underbrace{\mathbb{E}_{\rho}[L_{ce}(\boldsymbol{\theta})]}_{\text{Gibbs Error: } \rho\text{-average individual models' error}}$$

PAC-Bayesian upper-bounds (Alquier et al. 2016, Germain et al. 2016, Masegosa 2020)

$$\underbrace{L_{ce}(\rho)}_{\text{Gen. Error of } \rho\text{-combined model}} \overbrace{\leq}^{\text{Jensen Inequality}} \underbrace{\mathbb{E}_{\rho}[L_{ce}(\boldsymbol{\theta})]}_{\text{Gibbs Error: } \rho\text{-average individual models' error}} \overbrace{\lesssim}^{\text{with high probability}} \overbrace{\underbrace{\mathbb{E}_{\rho}[\hat{L}_{ce}(\boldsymbol{\theta}, D)]}_{\text{Empirical Gibbs Error}} + \frac{KL(\rho, \pi)}{n} + \frac{cte}{n}}^{\text{PAC-Bayes bound}}$$

PAC-Bayesian upper-bounds (Alquier et al. 2016, Germain et al. 2016, Masegosa 2020)

$$\underbrace{L_{ce}(\rho)}_{\text{Gen. Error of } \rho\text{-combined model}} \overset{\text{Jensen Inequality}}{\leq} \underbrace{\mathbb{E}_{\rho}[L_{ce}(\boldsymbol{\theta})]}_{\text{Gibbs Error: } \rho\text{-average individual models' error}} \overset{\text{with high probability}}{\lesssim} \overbrace{\underbrace{\mathbb{E}_{\rho}[\hat{L}_{ce}(\boldsymbol{\theta}, D)]}_{\text{Empirical Gibbs Error}} + \frac{KL(\rho, \pi)}{n} + \frac{cte}{n}}^{\text{PAC-Bayes bound}}$$

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

PAC-Bayesian upper-bounds (Alquier et al. 2016, Germain et al. 2016, Masegosa 2020)

$$\underbrace{L_{ce}(\rho)}_{\text{Gen. Error of } \rho\text{-combined model}} \overset{\text{Jensen Inequality}}{\leq} \underbrace{\mathbb{E}_{\rho}[L_{ce}(\boldsymbol{\theta})]}_{\text{Gibbs Error: } \rho\text{-average individual models' error}} \overset{\text{with high probability}}{\lesssim} \overbrace{\underbrace{\mathbb{E}_{\rho}[\hat{L}_{ce}(\boldsymbol{\theta}, D)]}_{\text{Empirical Gibbs Error}} + \frac{KL(\rho, \pi)}{n} + \frac{cte}{n}}^{\text{PAC-Bayes bound}}$$

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

The Bayesian posterior is a proxy

$$p(\boldsymbol{\theta}|D) \approx \arg\min_{\rho} L_{ce}(\rho)$$

Is the Bayesian approach an optimal distribution for model combination?

PAC-Bayesian upper-bounds (Alquier et al. 2016, Germain et al. 2016, Masegosa 2020)

$$\underbrace{L_{ce}(\rho)}_{\text{Gen. Error of } \rho\text{-combined model}} \overbrace{\leq}^{\text{Jensen Inequality}} \underbrace{\mathbb{E}_{\rho}[L_{ce}(\boldsymbol{\theta})]}_{\text{Gibbs Error: } \rho\text{-average individual models' error}} \overbrace{\lesssim}^{\text{with high probability}} \overbrace{\underbrace{\mathbb{E}_{\rho}[\hat{L}_{ce}(\boldsymbol{\theta}, D)]}_{\text{Empirical Gibbs Error}} + \frac{KL(\rho, \pi)}{n} + \frac{cte}{n}}^{\text{PAC-Bayes bound}}$$

1. The Bayesian posterior **converges (large sample limit) to** $\arg\min \mathbb{E}_{\rho}[L_{ce}(\boldsymbol{\theta})]$.

Is the Bayesian approach an optimal distribution for model combination?

PAC-Bayesian upper-bounds (Alquier et al. 2016, Germain et al. 2016, Masegosa 2020)

$$\underbrace{L_{ce}(\rho)}_{\substack{\textbf{Gen. Error of} \\ \rho\text{-combined model}}} \overbrace{\leq}^{\text{Jensen Inequality}} \underbrace{\mathbb{E}_{\rho}[L_{ce}(\boldsymbol{\theta})]}_{\substack{\textbf{Gibbs Error: } \rho\text{-average} \\ \text{individual models' error}}} \overbrace{\lesssim}^{\text{with high probability}} \overbrace{\underbrace{\mathbb{E}_{\rho}[\hat{L}_{ce}(\boldsymbol{\theta}, D)]}_{\textbf{Empirical Gibbs Error}} + \frac{KL(\rho, \pi)}{n} + \frac{cte}{n}}^{\text{PAC-Bayes bound}}$$

1. The Bayesian posterior **converges (large sample limit) to** $\arg\min \mathbb{E}_{\rho}[L_{ce}(\boldsymbol{\theta})]$.
2. The minimum of $\mathbb{E}_{\rho}[L(\boldsymbol{\theta})]$ is a **Dirac-delta distribution** centered around the **best possible single model $\boldsymbol{\theta}^{\star}$**.

$$\underbrace{\delta(\boldsymbol{\theta} - \boldsymbol{\theta}^{\star})}_{\text{Dirac-Delta distribution}} = \arg\min_{\rho} \mathbb{E}_{\rho}[L_{ce}(\boldsymbol{\theta})]$$

Is the Bayesian approach an optimal distribution for model combination?

PAC-Bayesian upper-bounds (Alquier et al. 2016, Germain et al. 2016, Masegosa 2020)

$$\underbrace{L_{ce}(\rho)}_{\substack{\textbf{Gen. Error of}\\ \rho\text{-combined model}}} \overbrace{\leq}^{\text{Jensen Inequality}} \underbrace{\mathbb{E}_{\rho}[L_{ce}(\boldsymbol{\theta})]}_{\substack{\textbf{Gibbs Error: } \rho\text{-average}\\ \text{individual models' error}}} \overbrace{\lesssim}^{\text{with high probability}} \overbrace{\underbrace{\mathbb{E}_{\rho}[\hat{L}_{ce}(\boldsymbol{\theta}, D)]}_{\textbf{Empirical Gibbs Error}} + \frac{KL(\rho, \pi)}{n} + \frac{cte}{n}}^{\text{PAC-Bayes bound}}$$

1. The Bayesian posterior **converges (large sample limit) to** $\arg\min \mathbb{E}_{\rho}[L_{ce}(\boldsymbol{\theta})]$.
2. The minimum of $\mathbb{E}_{\rho}[L(\boldsymbol{\theta})]$ is a **Dirac-delta distribution** centered around the **best possible single model $\boldsymbol{\theta}^{\star}$**.

$$\underbrace{\delta(\boldsymbol{\theta} - \boldsymbol{\theta}^{\star})}_{\text{Dirac-Delta distribution}} = \arg\min_{\rho} \mathbb{E}_{\rho}[L_{ce}(\boldsymbol{\theta})]$$

Is this solution optimal for model combination? (Masegosa 2020)

$$\delta(\boldsymbol{\theta} - \boldsymbol{\theta}^{\star}) = \arg\min_{\rho} L_{ce}(\rho)$$

Is the Bayesian approach an optimal distribution for model combination?

PAC-Bayesian upper-bounds (Alquier et al. 2016, Germain et al. 2016, Masegosa 2020)

$$\underbrace{L_{ce}(\rho)}_{\text{Gen. Error of } \rho\text{-combined model}} \overset{\text{Jensen Inequality}}{\leq} \underbrace{\mathbb{E}_{\rho}[L_{ce}(\boldsymbol{\theta})]}_{\text{Gibbs Error: } \rho\text{-average individual models' error}} \overset{\text{with high probability}}{\lesssim} \overbrace{\underbrace{\mathbb{E}_{\rho}[\hat{L}_{ce}(\boldsymbol{\theta}, D)]}_{\text{Empirical Gibbs Error}} + \frac{KL(\rho, \pi)}{n} + \frac{cte}{n}}^{\text{PAC-Bayes bound}}$$

1. The Bayesian posterior **converges (large sample limit) to** $\arg\min \mathbb{E}_{\rho}[L_{ce}(\boldsymbol{\theta})]$.
2. The minimum of $\mathbb{E}_{\rho}[L(\boldsymbol{\theta})]$ is a **Dirac-delta distribution** centered around the **best possible single model $\boldsymbol{\theta}^{\star}$**.

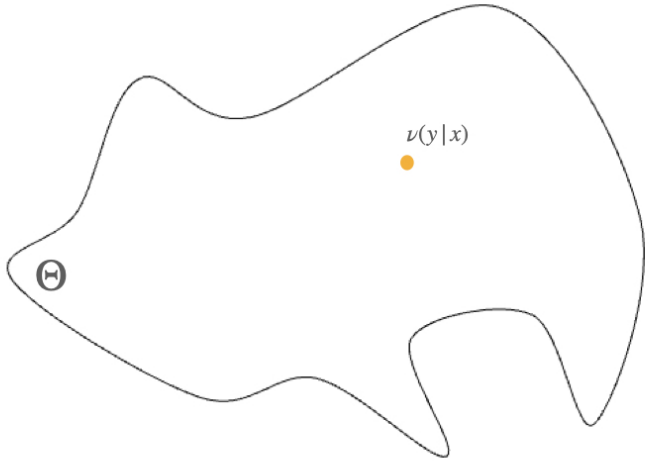
$$\underbrace{\delta(\boldsymbol{\theta} - \boldsymbol{\theta}^{\star})}_{\text{Dirac-Delta distribution}} = \arg\min_{\rho} \mathbb{E}_{\rho}[L_{ce}(\boldsymbol{\theta})]$$

Is this solution optimal for model combination? (Masegosa 2020)

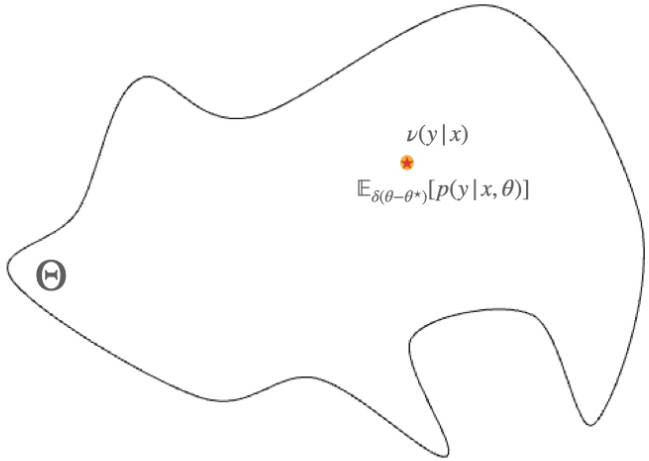
$$\delta(\boldsymbol{\theta} - \boldsymbol{\theta}^{\star}) = \arg\min_{\rho} L_{ce}(\rho)$$

Under **perfect model specification** (i.e. $\nu(y|\mathbf{x})$ belongs to the model class).

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

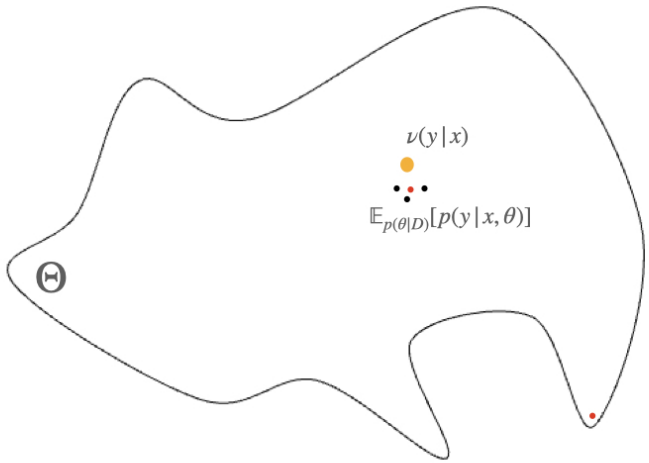


BMA **optimal under perfect specification** (Masegosa 2020)



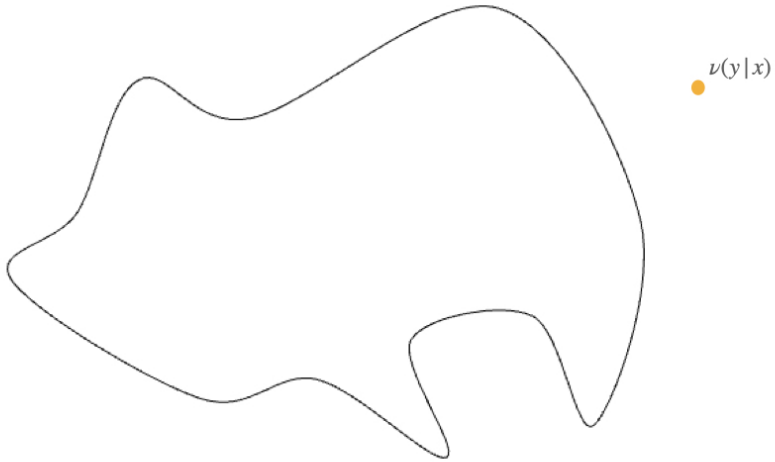
$$\underbrace{\delta(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)}_{\text{Dirac-Delta distribution}} = \arg\min_{\rho} \quad \underbrace{L_{ce}(\rho)}_{\textbf{Gen. Error of } \rho\text{-combined model}}$$

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

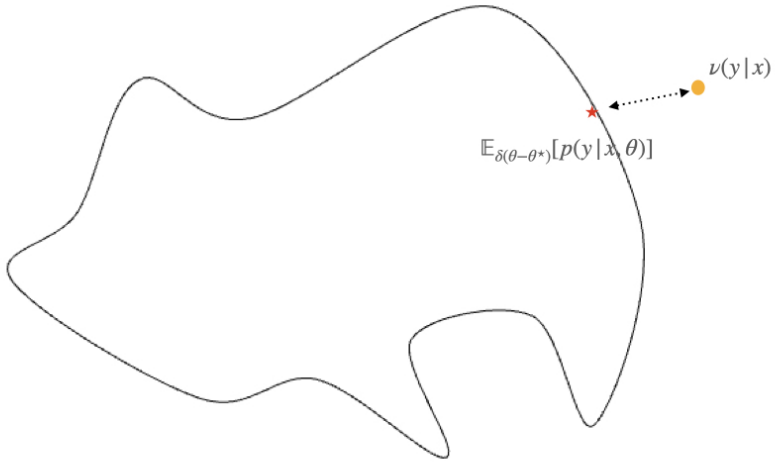


$$\underbrace{L_{ce}(\rho)}_{\substack{\textbf{Gen. Error} \text{ of} \\ \rho\text{-combined model}}} \overbrace{\leq}^{\substack{\text{Jensen} \\ \text{Inequality}}} \underbrace{\mathbb{E}_{\rho}[L_{ce}(\boldsymbol{\theta})]}_{\substack{\textbf{Gibbs Error}\text{: } \rho\text{-average} \\ \text{individual models' error}}} \overbrace{\lesssim}^{\substack{\text{with high} \\ \text{probability}}} \overbrace{\underbrace{\mathbb{E}_{\rho}[\hat{L}_{ce}(\boldsymbol{\theta}, D)]}_{\textbf{Empirical Gibbs Error}} + \frac{KL(\rho,\pi)}{n} + \frac{cte}{n}}^{\text{PAC-Bayes bound}}$$

BMA **optimal under model misspecification** (Masegosa 2020)

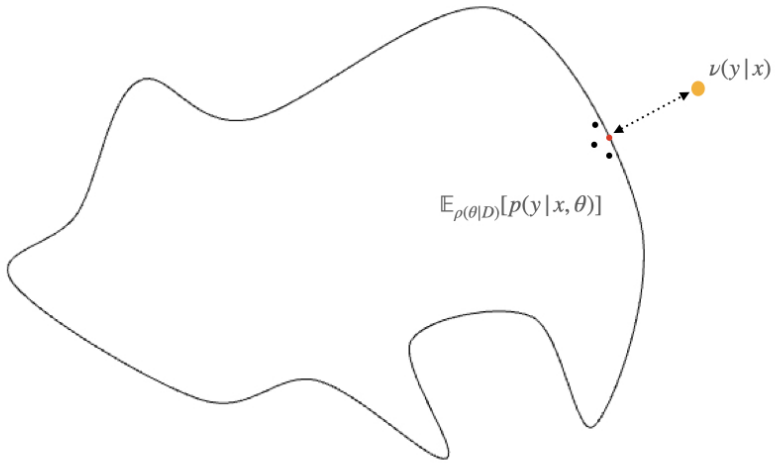


BMA **optimal under model misspecification** (Masegosa 2020)



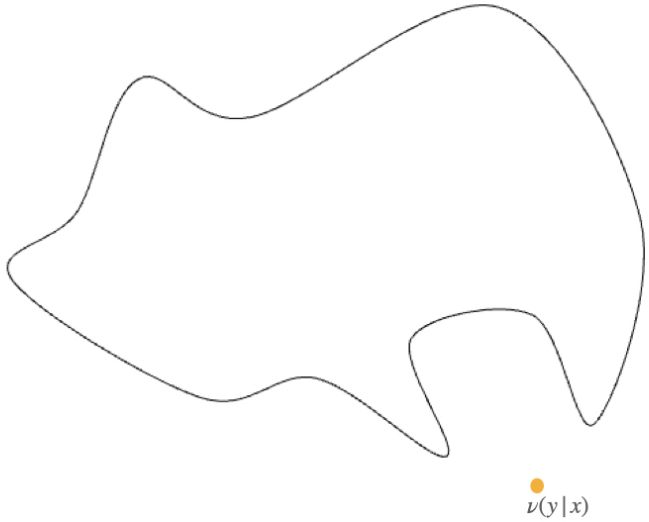
$$\underbrace{\delta(\boldsymbol{\theta}-\boldsymbol{\theta}^\star)}_{\text{Dirac-Delta distribution}} = \arg\min_{\rho} \quad \underbrace{L_{ce}(\rho)}_{\textbf{Gen. Error of } \rho\text{-combined model}}$$

BMA **optimal under model misspecification** (Masegosa 2020)

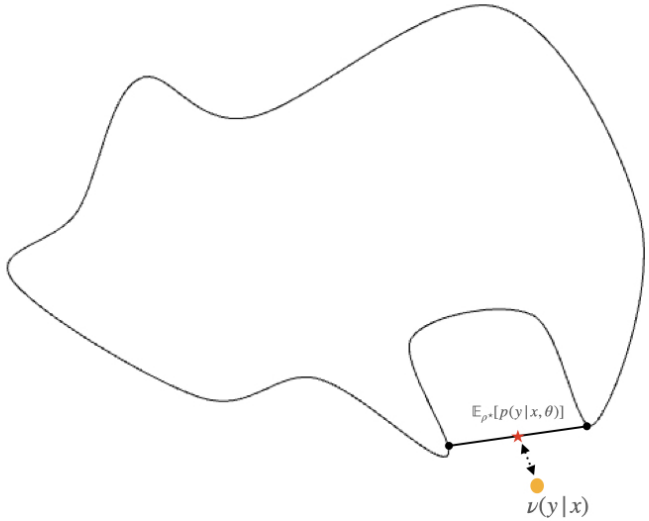


$$\underbrace{L_{ce}(\rho)}_{\text{Gen. Error of } \rho\text{-combined model}} \overbrace{\leq}^{\text{Jensen Inequality}} \underbrace{\mathbb{E}_{\rho}[L_{ce}(\boldsymbol{\theta})]}_{\text{Gibbs Error: } \rho\text{-average individual models' error}} \overbrace{\lesssim}^{\text{with high probability}} \overbrace{\underbrace{\mathbb{E}_{\rho}[\hat{L}_{ce}(\boldsymbol{\theta}, D)]}_{\text{Empirical Gibbs Error}} + \frac{KL(\rho,\pi)}{n} + \frac{cte}{n}}^{\text{PAC-Bayes bound}}$$

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

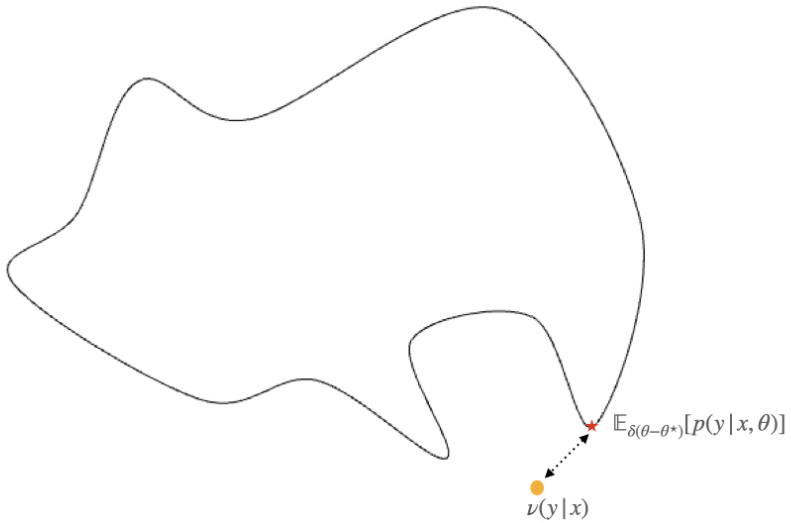


BMA **not optimal under model misspecification** (Masegosa 2020)



$$\underbrace{\delta(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)}_{\text{Dirac-Delta distribution}} \neq \arg\min_{\rho} \quad \underbrace{L_{ce}(\rho)}_{\text{**Gen. Error** of } \rho\text{-combined model}}$$

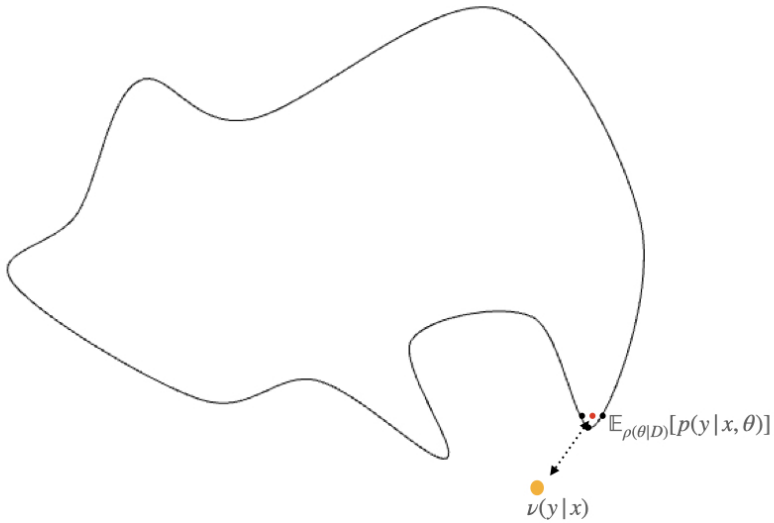
BMA **not optimal under model misspecification** (Masegosa 2020)



Jensen Inequality

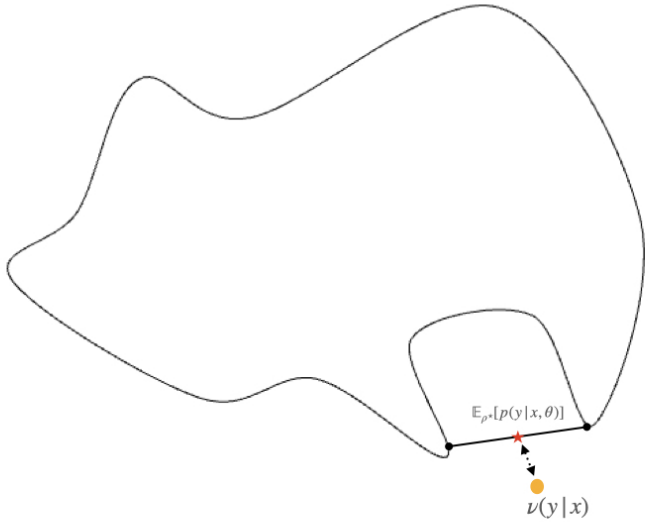
$$\underbrace{L_{ce}(\rho)}_{\text{Gen. Error of } \rho\text{-combined model}} \overbrace{\leq} \underbrace{\mathbb{E}_{\rho}[L_{ce}(\boldsymbol{\theta})]}_{\text{Gibbs Error: } \rho\text{-average individual models' error}}$$

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



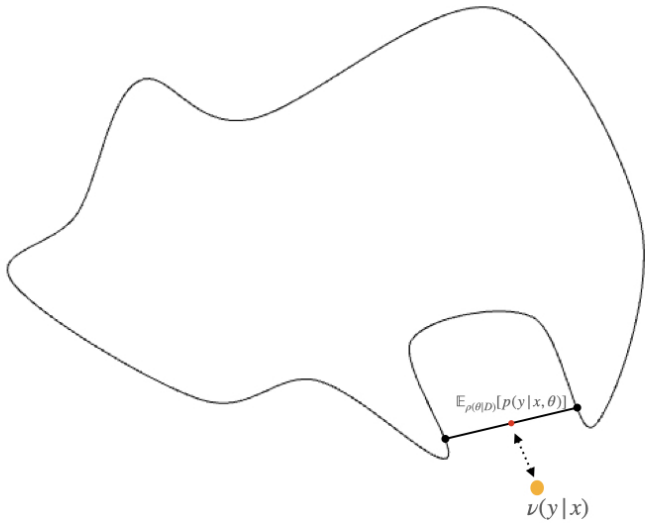
$$\underbrace{L_{ce}(\rho)}_{\substack{\textbf{Gen. Error} \text{ of} \\ \rho\text{-combined model}}} \overbrace{\leq}^{\substack{\text{Jensen} \\ \text{Inequality}}} \underbrace{\mathbb{E}_{\rho}[L_{ce}(\boldsymbol{\theta})]}_{\substack{\textbf{Gibbs Error}\text{: } \rho\text{-average} \\ \text{individual models' error}}} \overbrace{\lesssim}^{\substack{\text{with high} \\ \text{probability}}} \overbrace{\underbrace{\mathbb{E}_{\rho}[\hat{L}_{ce}(\boldsymbol{\theta}, D)]}_{\textbf{Empirical Gibbs Error}} + \frac{KL(\rho,\pi)}{n} + \frac{cte}{n}}^{\text{PAC-Bayes bound}}$$

Second-Order PAC-Bayesian Bounds (Masegosa 2020, Ortega et al. 2021)



$$\underbrace{L_{ce}(\rho)}_{\substack{\textbf{Gen. Error of} \\ \rho\text{-combined model}}} \overbrace{\leq}^{\substack{\text{Second-order} \\ \text{Jensen} \\ \text{Inequality}}} \underbrace{\mathbb{E}_{\rho}[L_{ce}(\boldsymbol{\theta})]}_{\textbf{Gibbs Error}} - \underbrace{\mathbb{D}_{ce}(\rho)}_{\textbf{Diversity of } \rho}$$

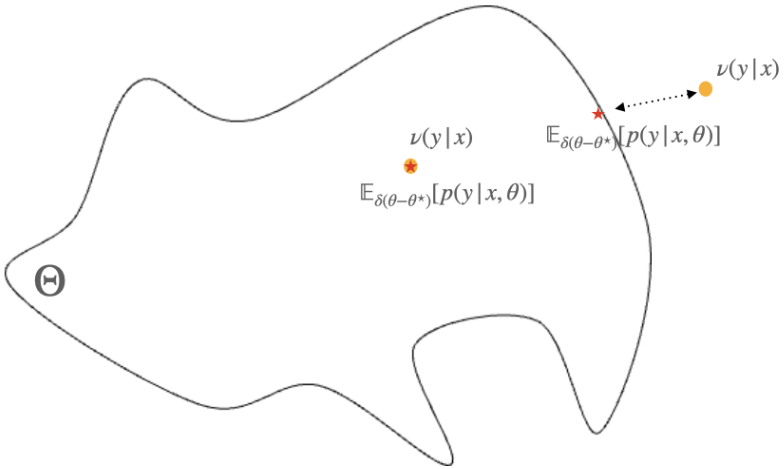
Second-Order PAC-Bayesian Bounds (Masegosa 2020)



$$\underbrace{\mathbb{E}_\rho[L_{ce}(\boldsymbol{\theta})]}_{\textbf{Gibbs Error}} - \underbrace{\mathbb{D}_{ce}(\rho)}_{\textbf{Diversity of } \rho} \overbrace{\lesssim}^{\text{with high probability}} \overbrace{\underbrace{\mathbb{E}_\rho[\hat{L}_{ce}(\boldsymbol{\theta}, D)]}_{\textbf{Empirical Gibbs Error}} - \underbrace{\hat{\mathbb{D}}_{ce}(\rho, D)}_{\textbf{Empirical Diversity}} + \frac{KL(\rho, \pi)}{n} + \frac{cte}{n}}^{\text{Second-Order PAC-Bayes bound}}$$

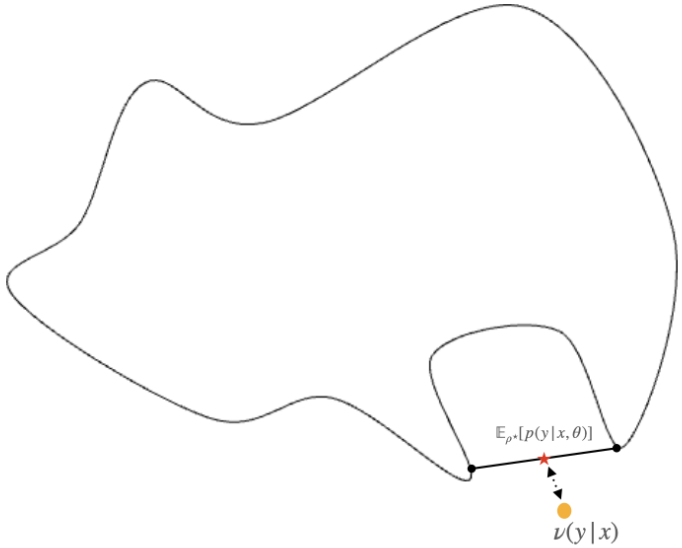
Two Regimes (Masegosa 2020, Ortega et al. 2021)

Bayesian Model Averaging



$$\underbrace{L(\boldsymbol{\theta}^\star)}_{\substack{\textbf{Gen. Error of} \\ \theta^\star \text{ model}}} < \underbrace{L_{ce}(\rho)}_{\substack{\textbf{Gen. Error of} \\ \rho\text{-combined model}}}$$

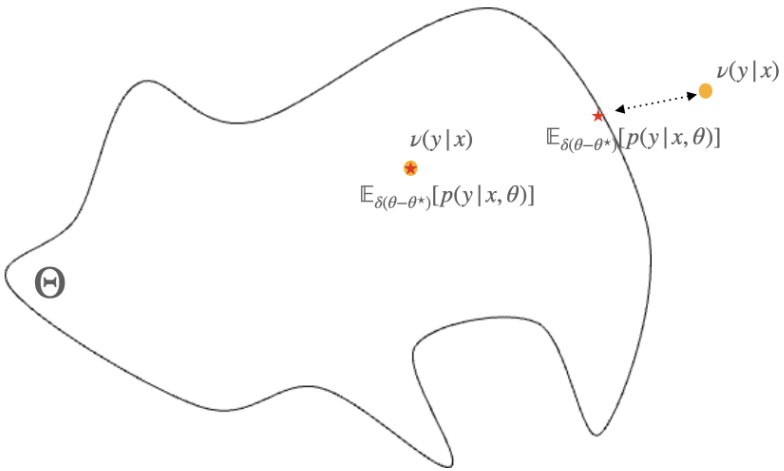
Model Combination



$$\underbrace{L(\boldsymbol{\theta}^\star)}_{\substack{\textbf{Gen. Error of} \\ \theta^\star \text{ model}}} > \underbrace{L_{ce}(\rho^\star)}_{\substack{\textbf{Gen. Error of} \\ \rho^\star\text{-combined model}}}$$

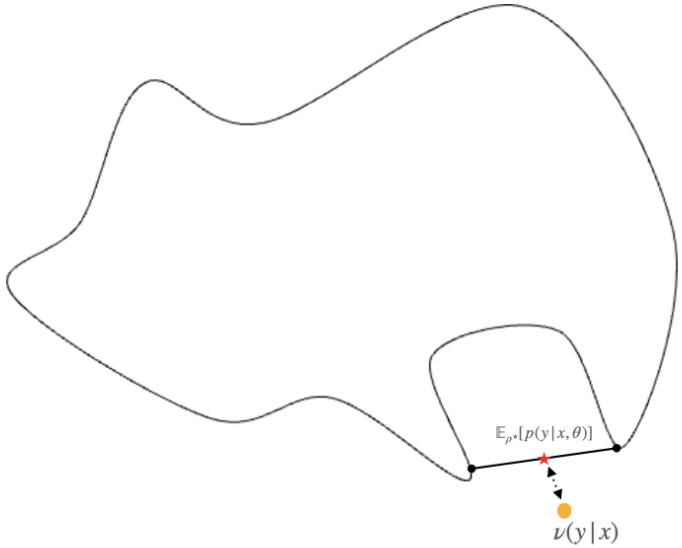
Two Regimes (Masegosa 2020, Ortega et al. 2021)

Bayesian Model Averaging



$$\underbrace{L(\boldsymbol{\theta}^\star)}_{\textbf{Gen. Error of } \theta^\star \text{ model}} < \underbrace{L_{ce}(\rho)}_{\textbf{Gen. Error of } \rho\text{-combined model}}$$

Model Combination

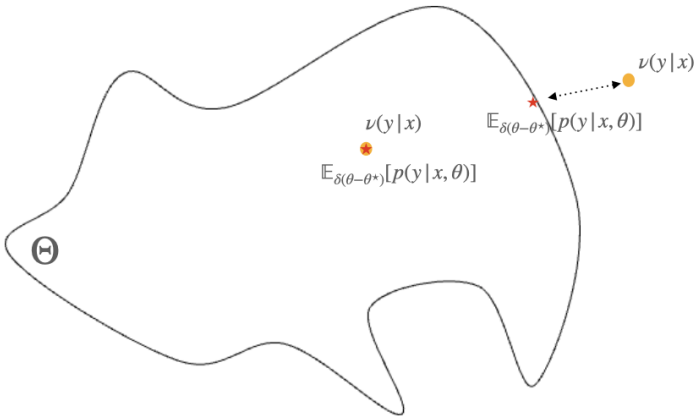


$$\underbrace{L(\boldsymbol{\theta}^\star)}_{\textbf{Gen. Error of } \theta^\star \text{ model}} > \underbrace{L_{ce}(\rho^\star)}_{\textbf{Gen. Error of } \rho^\star\text{-combined model}}$$

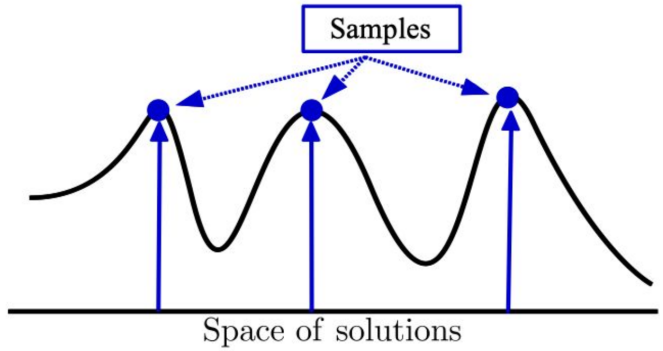
Which is the regime of Neural Networks Ensembles?

If Deep Ensembles work in the BMA regime

Bayesian Model Averaging Regime



Multi-Mode Posterior

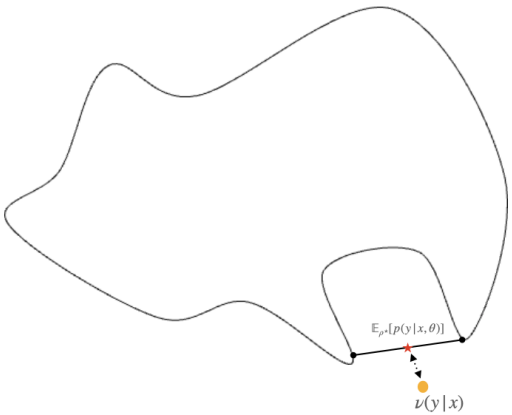


$$\underbrace{L(\boldsymbol{\theta}^\star)}_{\textbf{Gen. Error of } \theta^\star \text{ model}} < \underbrace{L_{ce}(p(\theta|D))}_{\textbf{Gen. Error of Bayesian Model Averaging}}$$

$$\underbrace{\mathbb{E}_{p(\theta|D)}[p(y \mid \mathbf{x}, \boldsymbol{\theta})]}_{\text{Bayesian Predictive posterior}} \approx \underbrace{\frac{1}{K}\sum_i \overbrace{p(y \mid \mathbf{x}, \boldsymbol{\theta}_i)}^{\text{Model Prediction}}}_{\text{Finite Sample Approximation}}$$

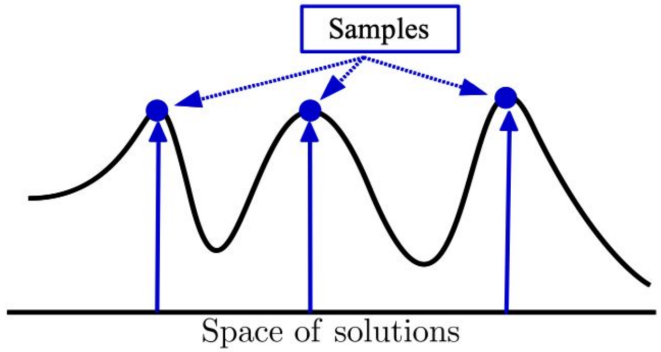
If Deep Ensembles work in the Model Combination regime

Model Combination Regime



$$\underbrace{L(\boldsymbol{\theta}^\star)}_{\substack{\textbf{Gen. Error of} \\ \theta^\star \text{ model}}} > \underbrace{L_{ce}(\rho^\star)}_{\substack{\textbf{Gen. Error of} \\ \rho^\star\text{-Combined Model}}}$$

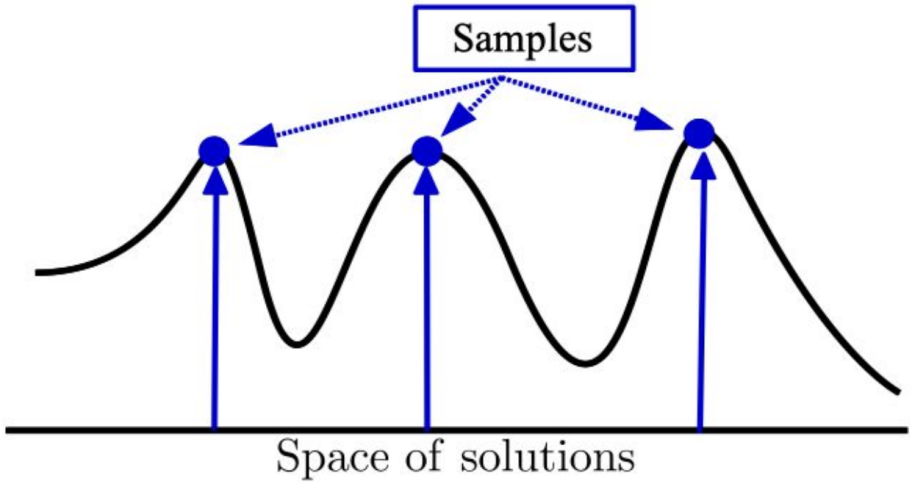
Each Local Optima is a **Different** Model



$$\underbrace{\mathbb{E}_{\rho^\star(\theta)}[p(y \mid \mathbf{x}, \boldsymbol{\theta})]}_{\rho^\star\text{-Combined Model}} \approx \underbrace{\frac{1}{K}\sum_i \overbrace{p(y \mid \mathbf{x}, \boldsymbol{\theta}_i)}^{\text{Model Prediction}}}_{\text{Uniform Model Combination}}$$

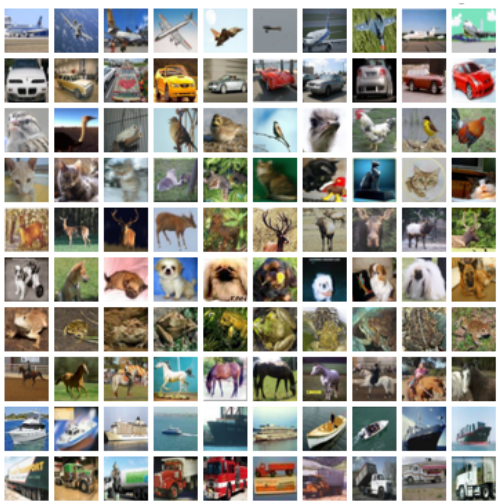
If Deep Ensembles work in the Model Combination regime

Each Local Optima is a **Different** Model

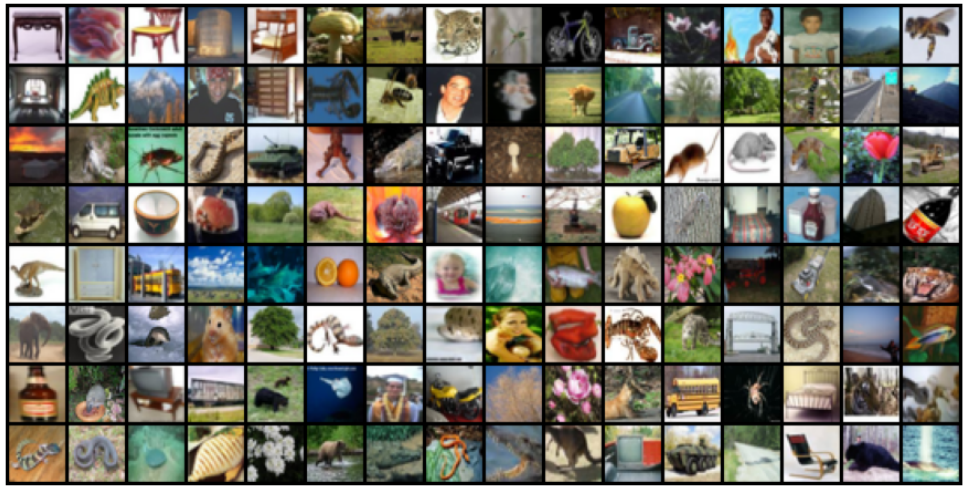


$$\underbrace{L_{ce}(\rho)}_{\substack{\textbf{Gen. Error of} \\ \rho\text{-combined model}}} \overbrace{\leq}^{\substack{\text{Second-order} \\ \text{Jensen} \\ \text{Inequality}}} \underbrace{\mathbb{E}_{\rho}[L_{ce}(\boldsymbol{\theta})]}_{\textbf{Gibbs Error}} - \underbrace{\mathbb{D}_{ce}(\rho)}_{\textbf{Diversity of } \rho}$$

Empirical Evidence (Ortega et al. 2021)



CIFAR 10

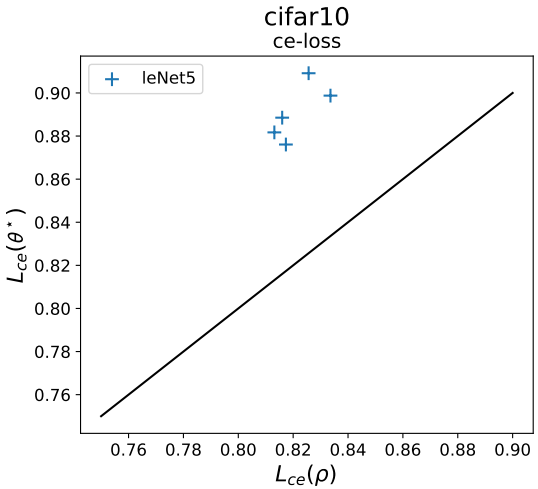


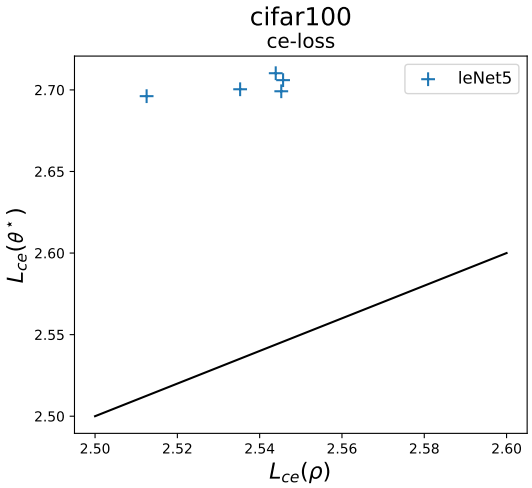
CIFAR 100

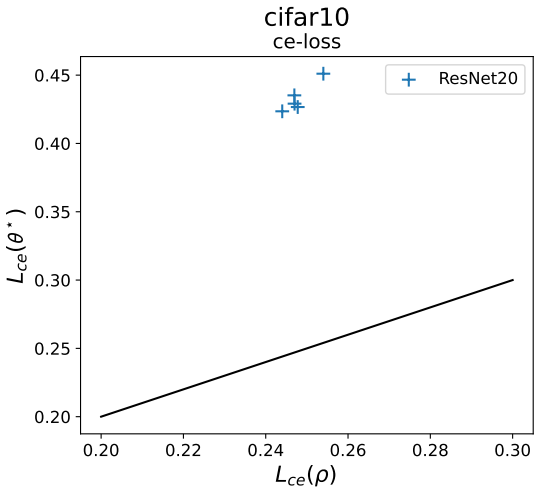
Experimental Settings

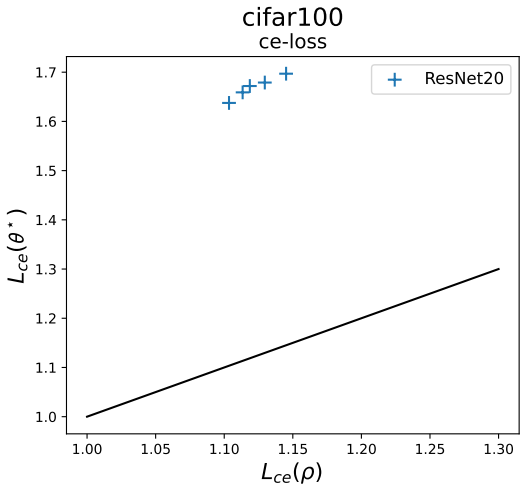
- Ensembles of **four** ResNet20/leNet5 Neural Networks.
- Each model was **independently run until convergence** (random initialization).
- Five repetitions.

Empirical Evidence (Ortega et al. 2021)





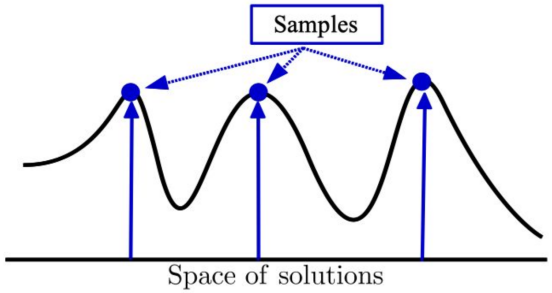




- Model Combination Regime **above** the black-line.
- Bayesian Model Averaging Regime **below** the black-line.

If Deep Ensembles work in the Model Combination regime

Each Local Optima is a **Different** Model



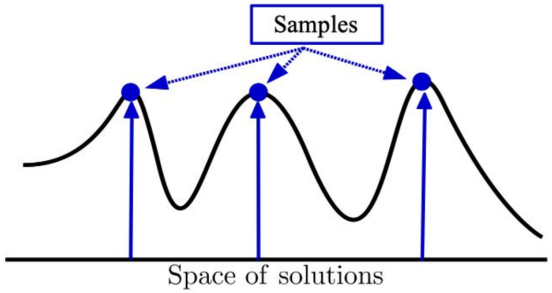
Corollary 5 of (Ortega et al. 2021)

If the **diversity of the ensemble is large enough**:

$$\underbrace{\mathbb{E}_{\rho}[L_{ce}(\boldsymbol{\theta})]}_{\textbf{Gibbs Error}} - \underbrace{L_{ce}(\boldsymbol{\theta}^{\star})}_{\textbf{Gen. Error of model } \theta^{\star}} \leq \underbrace{\mathbb{D}_{ce}(\rho)}_{\textbf{Diversity of } \rho}$$

If Deep Ensembles work in the Model Combination regime

Each Local Optima is a **Different** Model



Corollary 5 of (Ortega et al. 2021)

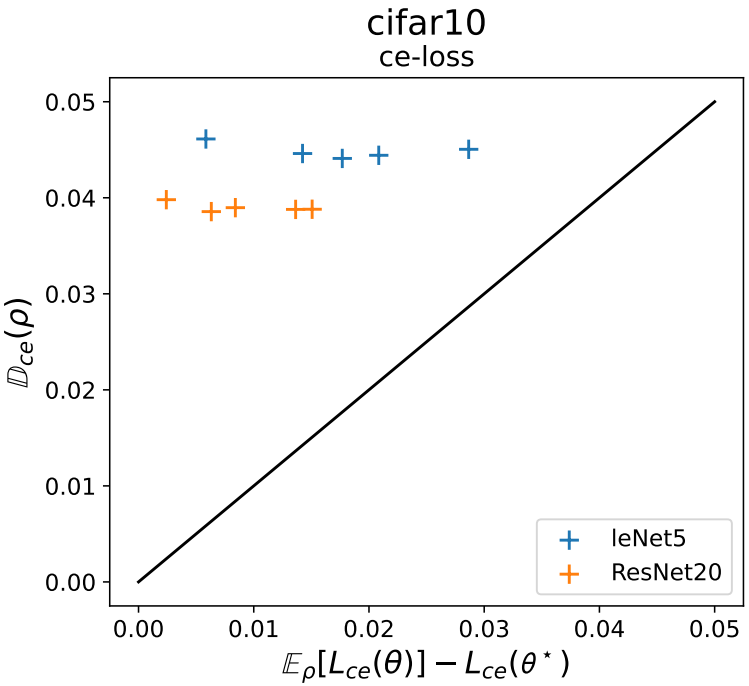
If the **diversity of the ensemble is large enough**:

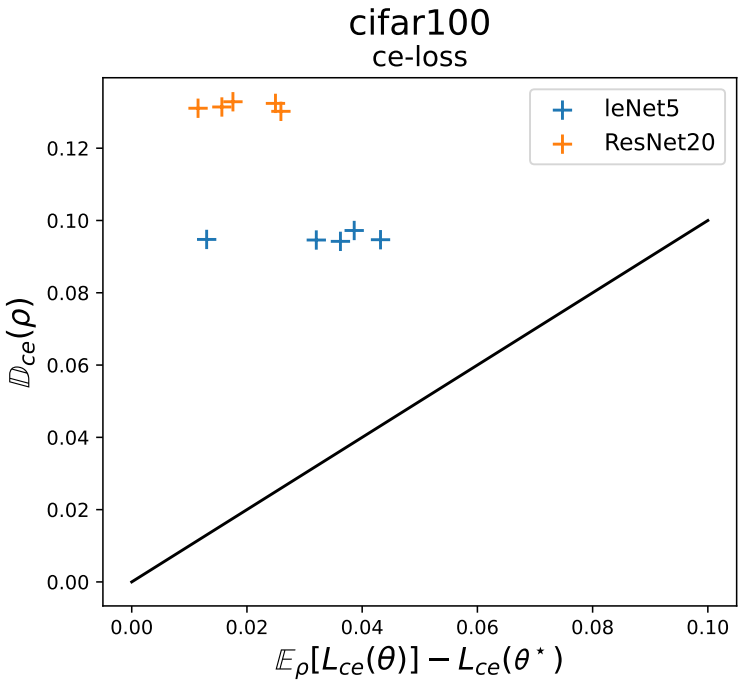
$$\underbrace{\mathbb{E}_{\rho}[L_{ce}(\boldsymbol{\theta})]}_{\textbf{Gibbs Error}} - \underbrace{L_{ce}(\boldsymbol{\theta}^{\star})}_{\textbf{Gen. Error of model } \theta^{\star}} \leq \underbrace{\mathbb{D}_{ce}(\rho)}_{\textbf{Diversity of } \rho}$$

then, the **generalization of the ensemble is better**:

$$\underbrace{L_{ce}(\boldsymbol{\theta}^{\star})}_{\textbf{Gen. Error of model } \theta^{\star}} > \underbrace{L_{ce}(\rho)}_{\textbf{Gen. Error of } \rho\text{-combined model}}$$

Empirical Evidence (Ortega et al. 2021)





- Model Combination Regime **above** the black-line.
- Bayesian Model Averaging Regime **below** the black-line.

Conclusions

- This analysis corroborates the notions of Thomas Minka:
  - BMA only tries to identify the best single model.
  - Model Combination exploit an enrich model space.

Conclusions

- This analysis corroborates the notions of Thomas Minka:
  - BMA only tries to identify the best single model.
  - Model Combination exploit an enrich model space.

- We have two regimes:
  - **BMA regime** is optimal when **your model class is not wrong**.
  - **MC regime** is optimal when **your model class is wrong**.

Conclusions

- This analysis corroborates the notions of Thomas Minka:
  - BMA only tries to identify the best single model.
  - Model Combination exploit an enrich model space.

- We have two regimes:
  - **BMA regime** is optimal when **your model class is not wrong**.
  - **MC regime** is optimal when **your model class is wrong**.

- Evidence supporting that Deep Ensembles works on the **model combination regime**.

Conclusions

- This analysis corroborates the notions of Thomas Minka:
  - BMA only tries to identify the best single model.
  - Model Combination exploit an enrich model space.

- We have two regimes:
  - **BMA regime** is optimal when **your model class is not wrong**.
  - **MC regime** is optimal when **your model class is wrong**.

- Evidence supporting that Deep Ensembles works on the **model combination regime**.

Ensembles are great when **Your model is Wrong!!**