Diversity and Generalization in Neural Network Ensembles

Andrés R. Masegosa

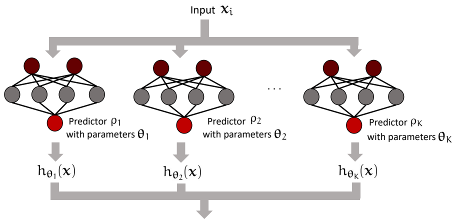
June 12, 2023

Aalborg University (Copenhagen Campus)

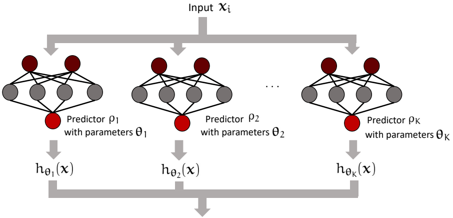
Ortega, L.A., Cabañas, R. and Masegosa, A. R., Diversity and Generalization in Neural Network Ensembles. AISTATS 2022.

Introduction and Motivation

Introduction: Ensembles of Neural Networks

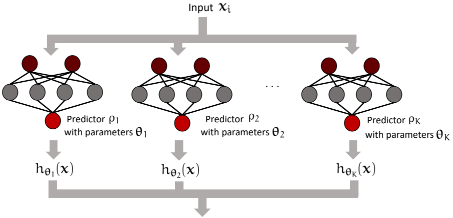


Introduction: Ensembles of Neural Networks



- Ensembles of NNs are recently getting a lot of attention.
  - Provide better **uncertainty quantification**.
  - More robust to **Out-Distribution-Data**.
  - Key properties in many real-world applications.

Introduction: Ensembles of Neural Networks



- Ensembles of NNs are recently getting a lot of attention.
  - Provide better **uncertainty quantification**.
  - More robust to **Out-Distribution-Data**.
  - Key properties in many real-world applications.

- Ongoing debate of **why ensembles of NNs work** so well:
  - **Ensemble's diversity** is widely used to justify ensemble performance.

Introduction: What is ensemble's diversity?

- Ensemble's **diversity** is a broad concept:
  - Ensemble's performance **jointly depends of the individual model's performance and the diversity among them**.

Introduction: What is ensemble's diversity?

- Ensemble's **diversity** is a broad concept:
  - Ensemble's performance **jointly depends of the individual model's performance and the diversity among them**.
  - An ensemble has **null diversity** if the predictions of individual models coincide on all the samples.

Introduction: What is ensemble's diversity?

- Ensemble's **diversity** is a broad concept:
  - Ensemble's performance **jointly depends of the individual model's performance and the diversity among them**.
  - An ensemble has **null diversity** if the predictions of individual models coincide on all the samples.
  - No advantage of having an ensemble when diversity is null.

Introduction: What is ensemble's diversity?

- Ensemble's **diversity** is a broad concept:
  - Ensemble's performance **jointly depends of the individual model's performance and the diversity among them**.
  - An ensemble has **null diversity** if the predictions of individual models coincide on all the samples.
  - No advantage of having an ensemble when diversity is null.

- Theoretically, diversity is **not a well-established concept**:
  - Different names: ambiguity, disagreement, etc.
  - Many different proposals to **define diversity**.
  - No theoretical analysis covering different different ensembles.

Our Contributions

- We built on previously published results:
  - **(Krogh and Vedelsby, 1994)**: Ensemble of regression models.
  - **(Masegosa, 2020)**: Bayesian model averaging.
  - **(Masegosa et al., 2020)**: Weighted Majority Vote.

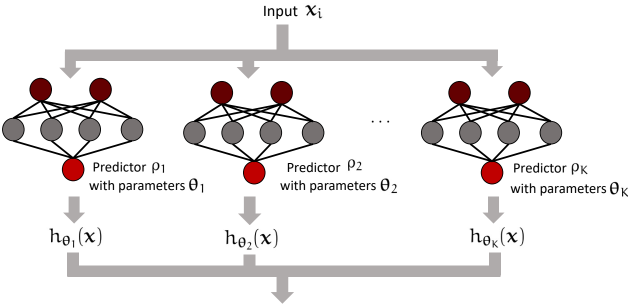
Our Contributions

- We built on previously published results:
  - **(Krogh and Vedelsby, 1994)**: Ensemble of regression models.
  - **(Masegosa, 2020)**: Bayesian model averaging.
  - **(Masegosa et al., 2020)**: Weighted Majority Vote.
- We introduce a theoretical framework to answer these questions:
  1) How to measure the diversity of an ensemble?.
  2) How is diversity related to the ensemble's generalization performance?.
  3) How can diversity be promoted by ensemble learning algorithms?.
- We derive a **common framework** for different types of ensembles.

Previous Knowledge

Basics on NNs ensembles

An ensemble trained with $D = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$ is the combination of different predictors.



Different Ensemble Methods

Regression Ensemble: Multiple regression models.

- Weighted Model Averaging: $MA_{\rho}(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{\theta}\sim\rho}[h_{\boldsymbol{\theta}}(\boldsymbol{x})]$.

Different Ensemble Methods

Regression Ensemble: Multiple regression models.

- Weighted Model Averaging: $MA_\rho(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{\theta}\sim\rho}[h_{\boldsymbol{\theta}}(\boldsymbol{x})]$.
- Squared loss :
  $L_{\mathsf{sq}}(\boldsymbol{\theta}) = \mathbb{E}_\nu[(h_{\boldsymbol{\theta}}(\boldsymbol{x}) - y)^2] \quad L_{\mathsf{sq}}(\rho) = \mathbb{E}_\nu[(MA_\rho(\boldsymbol{x}) - y)^2]$

Different Ensemble Methods

Regression Ensemble: Multiple regression models.

- Weighted Model Averaging: $MA_\rho(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{\theta}\sim\rho}[h_{\boldsymbol{\theta}}(\boldsymbol{x})]$.
- Squared loss :
  $L_{\mathrm{sq}}(\boldsymbol{\theta}) = \mathbb{E}_\nu[(h_{\boldsymbol{\theta}}(\boldsymbol{x}) - y)^2] \quad L_{\mathrm{sq}}(\rho) = \mathbb{E}_\nu[(MA_\rho(\boldsymbol{x}) - y)^2]$

Probabilistic Ensemble: Multiple probabilistic classification models.

- Weighted Model Averaging

Different Ensemble Methods

Regression Ensemble: Multiple regression models.

- Weighted Model Averaging: $MA_\rho(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{\theta}\sim\rho}[h_{\boldsymbol{\theta}}(\boldsymbol{x})]$.
- Squared loss :
  $L_{\mathrm{sq}}(\boldsymbol{\theta}) = \mathbb{E}_\nu[(h_{\boldsymbol{\theta}}(\boldsymbol{x}) - y)^2] \quad L_{\mathrm{sq}}(\rho) = \mathbb{E}_\nu[(MA_\rho(\boldsymbol{x}) - y)^2]$

Probabilistic Ensemble: Multiple probabilistic classification models.

- Weighted Model Averaging
- Cross-entropy loss

Different Ensemble Methods

Regression Ensemble: Multiple regression models.

- Weighted Model Averaging: $MA_\rho(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{\theta}\sim\rho}[h_{\boldsymbol{\theta}}(\boldsymbol{x})]$.
- Squared loss :
  $L_{\text{sq}}(\boldsymbol{\theta}) = \mathbb{E}_\nu[(h_{\boldsymbol{\theta}}(\boldsymbol{x}) - y)^2]$ $\quad L_{\text{sq}}(\rho) = \mathbb{E}_\nu[(MA_\rho(\boldsymbol{x}) - y)^2]$

Probabilistic Ensemble: Multiple probabilistic classification models.

- Weighted Model Averaging
- Cross-entropy loss

Majority Vote Ensemble: Multiple classification models.

- Weighted Majority Vote

Different Ensemble Methods

Regression Ensemble: Multiple regression models.

- Weighted Model Averaging: $MA_\rho(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{\theta}\sim\rho}[h_{\boldsymbol{\theta}}(\boldsymbol{x})]$.
- Squared loss :
  $L_{\mathsf{sq}}(\boldsymbol{\theta}) = \mathbb{E}_\nu[(h_{\boldsymbol{\theta}}(\boldsymbol{x}) - y)^2] \quad L_{\mathsf{sq}}(\rho) = \mathbb{E}_\nu[(MA_\rho(\boldsymbol{x}) - y)^2]$

Probabilistic Ensemble: Multiple probabilistic classification models.

- Weighted Model Averaging
- Cross-entropy loss

Majority Vote Ensemble: Multiple classification models.

- Weighted Majority Vote
- Zero-one loss

Diversity and Ensembles' Performance

Diversity and Generalization

Theorem 1

General Upper-bound for all the ensembles considered in this work:

$$\underbrace{L(\rho)}_{\substack{\text{Ensemble's} \\ \text{Expected Loss}}} \leq \alpha \left(\underbrace{\mathbb{E}_\rho[L(\boldsymbol{\theta})]}_{\substack{\text{Individual Models'} \\ \text{Expected Loss}}} - \underbrace{\mathbb{D}(\rho)}_{\substack{\text{Ensemble's} \\ \text{Diversity}}} \right)$$

where $\alpha = 4$ for the $0/1$-loss, otherwise, $\alpha = 1$.

The diversity term depends on the considered **loss function**:

$$\mathbb{D}(\rho) = \mathbb{E}_\nu \Big[\mathbb{V}_\rho \left(f(y, \boldsymbol{x}; \boldsymbol{\theta}) \right) \Big].$$

Diversity and Generalization

Theorem 1

General Upper-bound for all the ensembles considered in this work:

$$\underbrace{L(\rho)}_{\substack{\text{Ensemble's}\\\text{Expected Loss}}} \leq \alpha \left(\underbrace{\mathbb{E}_\rho[L(\boldsymbol{\theta})]}_{\substack{\text{Individual Models'}\\\text{Expected Loss}}} - \underbrace{\mathbb{D}(\rho)}_{\substack{\text{Ensemble's}\\\text{Diversity}}} \right)$$

where $\alpha = 4$ for the 0/1-loss, otherwise, $\alpha = 1$.

The diversity term depends on the considered **loss function**:

$$\mathbb{D}(\rho) = \mathbb{E}_\nu \left[\mathbb{V}_\rho \left(f(y, \boldsymbol{x}; \boldsymbol{\theta}) \right) \right].$$

- **Ensemble's performance depends** on both individual models' performance and diversity.

How to measure diversity?

Regression Ensemble: Multiple regression models.

$$\mathbb{D}_{\mathsf{sq}}(\rho) \;=\; \mathbb{E}_{\nu}\Big[\mathbb{V}_{\rho}(h_{\boldsymbol{\theta}}(\boldsymbol{x}))\Big]$$

How to measure diversity?

Regression Ensemble: Multiple regression models.

$$\mathbb{D}_{\mathsf{sq}}(\rho) \quad = \quad \mathbb{E}_{\nu}\Big[\mathbb{V}_{\rho}(h_{\boldsymbol{\theta}}(\boldsymbol{x}))\Big]$$

Probabilistic Ensemble: Multiple probabilistic classification models.

$$\mathbb{D}_{\mathsf{ce}}(\rho) \quad = \quad \mathbb{E}_{\nu}\left[\mathbb{V}_{\rho}\left(\frac{p(y \mid \boldsymbol{x}, \boldsymbol{\theta})}{\sqrt{2}\max_{\boldsymbol{\theta}} p(y \mid \boldsymbol{x}, \boldsymbol{\theta})}\right)\right]$$

How to measure diversity?

Regression Ensemble: Multiple regression models.

$$\mathbb{D}_{\mathsf{sq}}(\rho) \quad = \quad \mathbb{E}_{\nu}\Big[\mathbb{V}_{\rho}(h_{\boldsymbol{\theta}}(\boldsymbol{x}))\Big]$$

Probabilistic Ensemble: Multiple probabilistic classification models.

$$\mathbb{D}_{\mathsf{ce}}(\rho) \quad = \quad \mathbb{E}_{\nu}\left[\mathbb{V}_{\rho}\left(\frac{p(y \mid \boldsymbol{x}, \boldsymbol{\theta})}{\sqrt{2}\max_{\boldsymbol{\theta}} p(y \mid \boldsymbol{x}, \boldsymbol{\theta})}\right)\right]$$

Majority Vote Ensemble: Multiple classification models.

$$\mathbb{D}_{0/1}(\rho) \quad = \quad \mathbb{E}_{\nu}\Big[\mathbb{V}_{\rho}\Big(\mathbb{1}(h_{\boldsymbol{\theta}}(\boldsymbol{x}) \neq y)\Big)\Big]$$

Is $\mathbb{D}(\rho)$ a diversity measure?

A General Measure of Diversity:

$$\mathbb{D}(\rho) = \mathbb{E}_\nu\left[\mathbb{V}_\rho\left(f(y, \boldsymbol{x}; \boldsymbol{\theta})\right)\right].$$

Lemma

i) $\mathbb{D}(\rho) \geq 0$

ii) *If all individual models provide the same predictions, then* $\mathbb{D}(\rho) = 0$.

iii) $0 \leq \mathbb{D}(\rho) \leq \mathbb{E}_\rho[L(\boldsymbol{\theta})]$.

iv) $\mathbb{D}(\rho)$ *is invariant to reparametrizations.*

How to measure diversity?

A General Measure of Diversity:

$$\mathbb{D}(\rho) = \mathbb{E}_{\nu}\Big[\mathbb{V}_{\rho}\left(f(y, \boldsymbol{x}; \boldsymbol{\theta})\right)\Big]$$

Theorem

The diversity term $\mathbb{D}(\rho)$ can be written as

$$\mathbb{D}(\rho) = \mathbb{V}_{\nu\times\rho}\Big(f(y, \boldsymbol{x}; \boldsymbol{\theta})\Big) - \mathbb{E}_{\rho\times\rho}\Big[Cov_{\nu}(f(y, \boldsymbol{x}; \boldsymbol{\theta}), f(y, \boldsymbol{x}; \boldsymbol{\theta}'))\Big]$$

where $Cov_{\nu}(\cdot,\cdot)$ is the co-variance between two models.

How to measure diversity?

A General Measure of Diversity:

$$\mathbb{D}(\rho) = \mathbb{E}_{\nu}\Big[\mathbb{V}_{\rho}\left(f(y, \boldsymbol{x}; \boldsymbol{\theta})\right)\Big]$$

Theorem

The diversity term $\mathbb{D}(\rho)$ can be written as

$$\mathbb{D}(\rho) = \mathbb{V}_{\nu\times\rho}\Big(f(y, \boldsymbol{x}; \boldsymbol{\theta})\Big) - \mathbb{E}_{\rho\times\rho}\Big[Cov_{\nu}(f(y, \boldsymbol{x}; \boldsymbol{\theta}), f(y, \boldsymbol{x}; \boldsymbol{\theta}'))\Big]$$

where $Cov_{\nu}(\cdot, \cdot)$ is the co-variance between two models.

- First term helps to explain why **randomized models** improve ensemble performance.

How to measure diversity?

A General Measure of Diversity:

$$\mathbb{D}(\rho) = \mathbb{E}_{\nu}\left[\mathbb{V}_{\rho}\left(f(y, \boldsymbol{x}; \boldsymbol{\theta})\right)\right]$$

Theorem

The diversity term $\mathbb{D}(\rho)$ can be written as

$$\mathbb{D}(\rho) = \mathbb{V}_{\nu\times\rho}\Big(f(y, \boldsymbol{x}; \boldsymbol{\theta})\Big) - \mathbb{E}_{\rho\times\rho}\Big[Cov_{\nu}(f(y, \boldsymbol{x}; \boldsymbol{\theta}), f(y, \boldsymbol{x}; \boldsymbol{\theta}'))\Big]$$

where $Cov_{\nu}(\cdot, \cdot)$ is the co-variance between two models*.*

- First term helps to explain why **randomized models** improve ensemble performance.
- Second term helps to explain why **independent and anti-correlated models** improve ensemble performance.

How is Diversity Related to the Performance of an Ensemble?

Diversity and Generalization

Theorem 1

General Upper-bound for all the ensembles considered in this work:

$$\underbrace{L(\rho)}_{\substack{\text{Ensemble's} \\ \text{Expected Loss}}} \leq \alpha \left(\underbrace{\mathbb{E}_\rho[L(\boldsymbol{\theta})]}_{\substack{\text{Individual Models'} \\ \text{Expected Loss}}} - \underbrace{\mathbb{D}(\rho)}_{\substack{\text{Ensemble's} \\ \text{Diversity}}} \right)$$

where $\alpha = 4$ for the $0/1$-loss, otherwise, $\alpha = 1$.

- **Ensemble's performance depends** on both individual models' performance and diversity among them.

How is Diversity Related to the Performance of an Ensemble?

A General Measure of Diversity:

$$\mathbb{D}(\rho) = \mathbb{E}_\nu\Big[\mathbb{V}_\rho\left(f(y, \boldsymbol{x}; \boldsymbol{\theta})\right)\Big].$$

Question: How much do we gain by ensembling a set of models wrt randomly choosing them?

How is Diversity Related to the Performance of an Ensemble?

A General Measure of Diversity:

$$\mathbb{D}(\rho) = \mathbb{E}_\nu\Big[\mathbb{V}_\rho\left(f(y, \boldsymbol{x}; \boldsymbol{\theta})\right)\Big].$$

Question: How much do we gain by ensembling a set of models wrt randomly choosing them?

Corollary

Under these settings, we have that

$$\mathbb{D}(\rho) \leq \underbrace{\mathbb{E}_\rho[L(\boldsymbol{\theta})] - \frac{1}{\alpha}L(\rho)}_{\textit{Ensemble's Gap}}$$

Answer: Larger diversity induces larger gains.

How is Diversity Related to the Performance of an Ensemble?

A General Measure of Diversity:

$$\mathbb{D}(\rho) = \mathbb{E}_{\nu}\Big[\mathbb{V}_{\rho}\left(f(y, \boldsymbol{x}; \boldsymbol{\theta})\right)\Big].$$

Question: When is an ensemble better that best individual model?

How is Diversity Related to the Performance of an Ensemble?

A General Measure of Diversity:

$$\mathbb{D}(\rho) = \mathbb{E}_\nu\Big[\mathbb{V}_\rho\left(f(y, \boldsymbol{x}; \boldsymbol{\theta})\right)\Big].$$

Question: When is an ensemble better that best individual model?

Corollary

$$\underbrace{\mathbb{E}_\rho[L(\boldsymbol{\theta})] - \frac{1}{\alpha}L(\boldsymbol{\theta}^\star)}_{\substack{\textit{Single Model's} \\ \textit{Error Gap}}} < \underbrace{\mathbb{D}(\rho)}_{\substack{\textit{Ensemble's} \\ \textit{Diversity}}} \implies \underbrace{L(\rho)}_{\substack{\textit{Ensemble's} \\ \textit{Expected Loss}}} < \underbrace{L(\boldsymbol{\theta}^\star)}_{\substack{\textit{Single Model's} \\ \textit{Expected Loss}}}.$$

Answer: If the diversity of the ensemble is large enough, then an ensemble outperforms the best single model.

How to Exploit Diversity to Learn Ensembles?

Diversity and Generalization

Theorem 1

General Upper-bound for all the ensembles considered in this work:

$$\underbrace{L(\rho)}_{\substack{\text{Ensemble's} \\ \text{Expected Loss}}} \leq \alpha \left(\underbrace{\mathbb{E}_\rho[L(\boldsymbol{\theta})]}_{\substack{\text{Individual Models'} \\ \text{Expected Loss}}} - \underbrace{\mathbb{D}(\rho)}_{\substack{\text{Ensemble's} \\ \text{Diversity}}} \right)$$

where $\alpha = 4$ for the $0/1$-loss, otherwise, $\alpha = 1$.

- This inequality depends on the **data generating distribution**.

How to Exploit Diversity to Learn Ensembles?

A PAC-Bayesian Bound

For distribution $\pi(\boldsymbol{\theta})$ independent of D, with probability at least $1-\delta$ over draws of training data $D \sim \nu^n(y, \boldsymbol{x})$ (*i.e.*, i.i.d.), for all $\lambda > 0$, for all distribution ρ over $\boldsymbol{\Theta}$, simultaneously,

$$\underbrace{L(\rho)}_{\substack{\text{Ensemble's}\\ \text{Expected Loss}}} \leq \alpha \left(\underbrace{\mathbb{E}_\rho[\hat{L}(\boldsymbol{\theta}, D)]}_{\substack{\text{Averaged}\\ \text{Empirical Loss}}} - \underbrace{\hat{\mathbb{D}}(\rho, D)}_{\substack{\text{Ensemble's}\\ \text{Empirical Diversity}}} + \underbrace{\frac{2KL(\rho \mid \pi)}{\lambda n}}_{\text{Regularization}} + \frac{\epsilon(\nu, \pi, \lambda, n, \delta)}{\lambda n} \right)$$

How to Exploit Diversity to Learn Ensembles?

A PAC-Bayesian Bound

For distribution $\pi(\boldsymbol{\theta})$ independent of D, with probability at least $1-\delta$ over draws of training data $D \sim \nu^n(y, \boldsymbol{x})$ (*i.e.*, i.i.d.), for all $\lambda > 0$, for all distribution ρ over $\boldsymbol{\Theta}$, simultaneously,

$$\underbrace{L(\rho)}_{\substack{\text{Ensemble's} \\ \text{Expected Loss}}} \leq \alpha \left(\underbrace{\mathbb{E}_\rho[\hat{L}(\boldsymbol{\theta}, D)]}_{\substack{\text{Averaged} \\ \text{Empirical Loss}}} - \underbrace{\hat{\mathbb{D}}(\rho, D)}_{\substack{\text{Ensemble's} \\ \text{Empirical Diversity}}} + \underbrace{\frac{2KL(\rho \mid \pi)}{\lambda n}}_{\text{Regularization}} + \frac{\epsilon(\nu, \pi, \lambda, n, \delta)}{\lambda n} \right)$$

- Find the ρ minimizing this PAC-Bayesian Bound.

How to Exploit Diversity to Learn Ensembles?

A PAC-Bayesian Bound

For distribution $\pi(\boldsymbol{\theta})$ independent of D, with probability at least $1-\delta$ over draws of training data $D \sim \nu^n(y, \boldsymbol{x})$ (*i.e.*, i.i.d.), for all $\lambda > 0$, for all distribution ρ over $\boldsymbol{\Theta}$, simultaneously,

$$\underbrace{L(\rho)}_{\substack{\text{Ensemble's} \\ \text{Expected Loss}}} \leq \alpha \left(\underbrace{\mathbb{E}_\rho[\hat{L}(\boldsymbol{\theta}, D)]}_{\substack{\text{Averaged} \\ \text{Empirical Loss}}} - \underbrace{\hat{\mathbb{D}}(\rho, D)}_{\substack{\text{Ensemble's} \\ \text{Empirical Diversity}}} + \underbrace{\frac{2KL(\rho \mid \pi)}{\lambda n}}_{\text{Regularization}} + \frac{\epsilon(\nu, \pi, \lambda, n, \delta)}{\lambda n} \right)$$

- Find the ρ minimizing this PAC-Bayesian Bound.
- We move to a continuous hypothesis space.

How to Exploit Diversity to Learn Ensembles?

Ensemble Learning algorithm as a mixture model

$$\rho(\boldsymbol{\theta}|\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K, \sigma^2) = \frac{1}{K}\sum_{k=1}^{K}\mathcal{N}(\boldsymbol{\theta};\ \boldsymbol{\theta}_k, \sigma^2 I).$$

How to Exploit Diversity to Learn Ensembles?

Ensemble Learning algorithm as a mixture model

$$\rho(\boldsymbol{\theta}|\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K, \sigma^2) = \frac{1}{K}\sum_{k=1}^{K}\mathcal{N}(\boldsymbol{\theta};\ \boldsymbol{\theta}_k, \sigma^2 I).$$

Learning Objective (P2B-Ensemble)

$$\min_{\theta_1,\ldots,\theta_K} \underbrace{\mathbb{E}_\rho[\hat{L}(\boldsymbol{\theta}, D)]}_{\substack{\text{Averaged} \\ \text{Empirical Loss}}} \ - \ \underbrace{\hat{\mathbb{D}}(\rho, D)}_{\substack{\text{Ensemble's} \\ \text{Empirical Diversity}}} \ - \ \underbrace{\frac{2\mathbb{E}_\rho[\ln \pi(\boldsymbol{\theta})]}{\lambda n}}_{\text{Regularization}}$$

Empirical Validation

Empirical validation: Experimental Settings

Tasks

- **Regression Task**: Wine-Quality dataset.
- **Classification Task**: Cifar10 and Cifar100 data sets.

Empirical validation: Experimental Settings

Tasks

- **Regression Task**: Wine-Quality dataset.
- **Classification Task**: Cifar10 and Cifar100 data sets.

Models

- **Regression Task**: MLP with 50 hidden units.
- **Classification Task**: LeNet5 and ResNet20 convolutional networks.

Empirical validation: Experimental Settings

Tasks

- **Regression Task**: Wine-Quality dataset.
- **Classification Task**: Cifar10 and Cifar100 data sets.

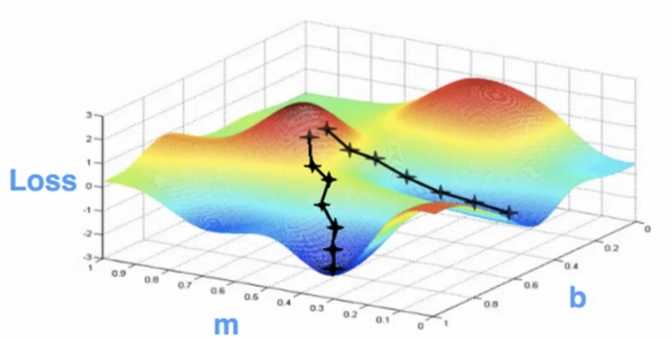
Models

- **Regression Task**: MLP with 50 hidden units.
- **Classification Task**: LeNet5 and ResNet20 convolutional networks.

Learning Algorithms

- **P2B-Ensemble**: K models jointly learned promoting diversity.
- **Ensemble**: K models independently learned.

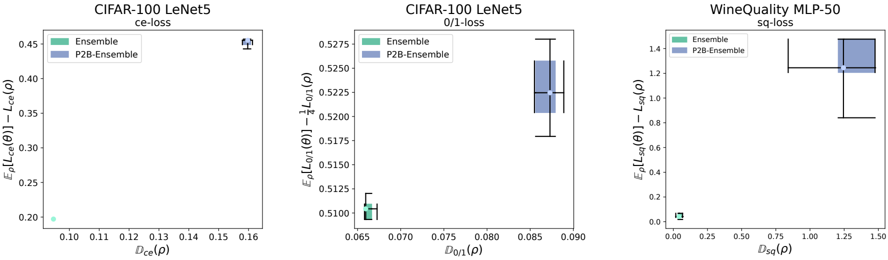
Ensemble Learning



(Manish Kumar, 2018)

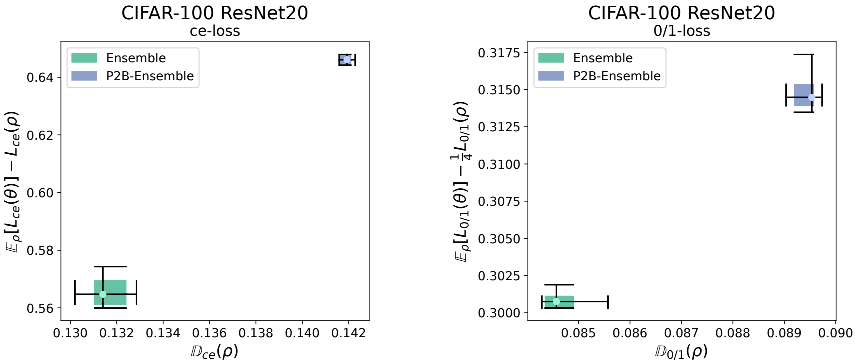
- Ensemble composed by K different local minima.

Empirical Validation



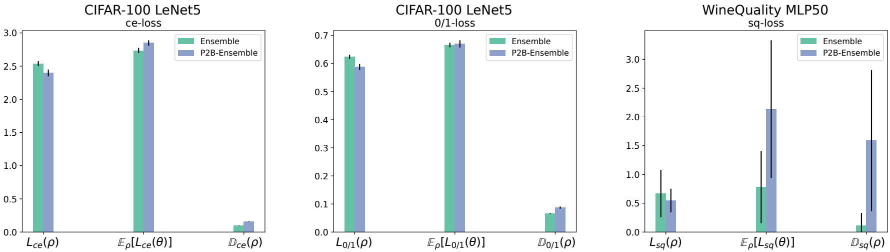
- Higher diversity correlates with higher gains by ensembling.
- Standard ensemble methods implicitly promote diversity.
- P2B-Ensemble finds ensembles with higher diversity.

Empirical validation



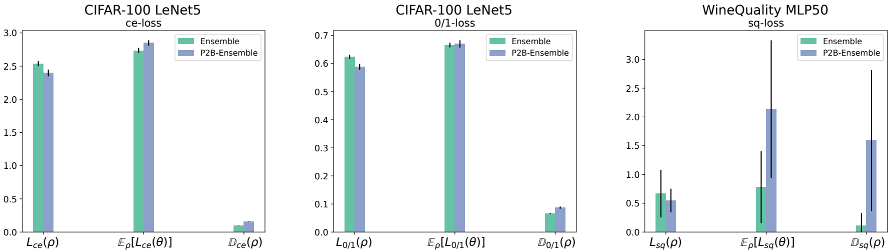
- Higher diversity correlates with higher gains by ensembling.
- Standard ensemble methods implicitly promote diversity.
- P2B-Ensemble finds ensembles with higher diversity.

Empirical validation



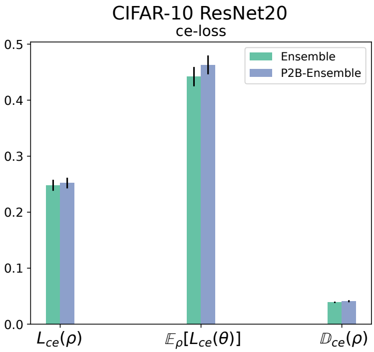
- Explicitly promoting diversity (ie. P2B-Ensemble) gives rise to better ensembles.

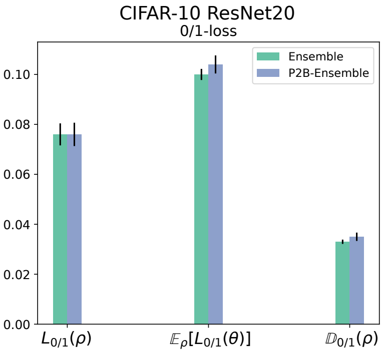
Empirical validation



- Explicitly promoting diversity (ie. P2B-Ensemble) gives rise to better ensembles.
- But not always...

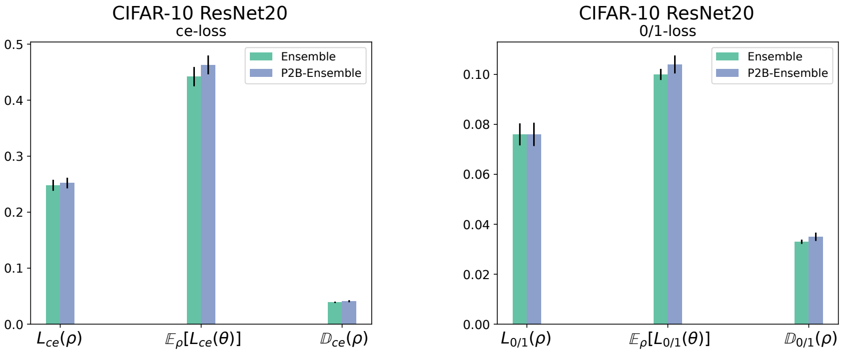
Empirical validation





- P2B-Ensemble is not able to learn better ensembles.

Empirical validation



- P2B-Ensemble is not able to learn better ensembles.
- Why?
    - Because big neural networks **works in the interpolation regime**.

How to Exploit Diversity to Learn Ensembles?

Learning Objective

$$\min_{\theta_1,\ldots,\theta_K} \underbrace{\mathbb{E}_\rho[\hat{L}(\boldsymbol{\theta}, D)]}_{\substack{\text{Averaged} \\ \text{Empirical Loss}}} - \underbrace{\hat{\mathbb{D}}(\rho, D)}_{\substack{\text{Ensemble's} \\ \text{Empirical Diversity}}} - \underbrace{\frac{2\mathbb{E}_\rho[\ln \pi(\boldsymbol{\theta})]}{\lambda n}}_{\text{Regularization}}$$

How to Exploit Diversity to Learn Ensembles?

Learning Objective

$$\min_{\theta_1,\ldots,\theta_K} \underbrace{\mathbb{E}_\rho[\hat{L}(\boldsymbol{\theta}, D)]}_{\substack{\text{Averaged} \\ \text{Empirical Loss}}} - \underbrace{\hat{\mathbb{D}}(\rho, D)}_{\substack{\text{Ensemble's} \\ \text{Empirical Diversity}}} - \underbrace{\frac{2\mathbb{E}_\rho[\ln \pi(\boldsymbol{\theta})]}{\lambda n}}_{\text{Regularization}}$$

Inequality

$$0 \leq \hat{\mathbb{D}}(\rho, D) \leq \mathbb{E}_\rho[\hat{L}(\boldsymbol{\theta}, D)]$$

How to Exploit Diversity to Learn Ensembles?

Learning Objective

$$\min_{\theta_1,\ldots,\theta_K} \underbrace{\mathbb{E}_\rho[\hat{L}(\boldsymbol{\theta}, D)]}_{\substack{\text{Averaged} \\ \text{Empirical Loss}}} - \underbrace{\hat{\mathbb{D}}(\rho, D)}_{\substack{\text{Ensemble's} \\ \text{Empirical Diversity}}} - \underbrace{\frac{2\mathbb{E}_\rho[\ln \pi(\boldsymbol{\theta})]}{\lambda n}}_{\text{Regularization}}$$

Inequality

$$0 \leq \hat{\mathbb{D}}(\rho, D) \leq \mathbb{E}_\rho[\hat{L}(\boldsymbol{\theta}, D)]$$

In the interpolation regime

$$\mathbb{E}_\rho[\hat{L}(\boldsymbol{\theta}, D)] \approx 0 \Rightarrow \hat{\mathbb{D}}(\rho, D) \approx 0$$

- The empirical diversity does not provide any signal to the gradient.

Conclusions and Future Work

Conclusions

- We can formally speak about ensemble's diversity.
- Applies to very different ensemble methods.
- Useful to understand and derive learning algorithms.

Conclusions and Future Work

Conclusions

- We can formally speak about ensemble's diversity.
- Applies to very different ensemble methods.
- Useful to understand and derive learning algorithms.

Limitations

- Diversity's linear dependency: not accurate in all cases (Germain et al. 2015, Wu et al. 2021)
- Only second-order interactions.
- Learning in the interpolation-regime.

Conclusions and Future Work

Conclusions

- We can formally speak about ensemble's diversity.
- Applies to very different ensemble methods.
- Useful to understand and derive learning algorithms.

Limitations

- Diversity's linear dependency: not accurate in all cases (Germain et al. 2015, Wu et al. 2021)
- Only second-order interactions.
- Learning in the interpolation-regime.

Future Works

- Promote diversity using a external (non-labelled) dataset.

Questions?

Andrés, L.A.O., Cabañas, R. and Masegosa, A. R., Diversity and Generalization in Neural Network Ensembles. AISTATS 2022.

Measuring Diversity

- For **regression ensembles**, (Krogh and Vedelsby, 1994) showed that:

$$\underbrace{L_{\mathsf{sq}}(\rho)}_{\substack{\text{Ensemble's}\\\text{Expected Loss}}} = \mathbb{E}_\rho[\underbrace{\mathbb{E}_\nu[(y - h_{\boldsymbol{\theta}}(\boldsymbol{x}))^2]}_{\substack{\text{Individual Models'}\\\text{Expected Loss}}}] - \mathbb{E}_\nu[\underbrace{\mathbb{E}_\rho[(h_{\boldsymbol{\theta}}(\boldsymbol{x}) - \mathbb{E}_\rho[h_{\boldsymbol{\theta}}(\boldsymbol{x})])^2]}_{\substack{\text{Variance among}\\\text{individual models}}}]$$

Measuring Diversity

- For **regression ensembles**, (Krogh and Vedelsby, 1994) showed that:

$$\underbrace{L_{\mathsf{sq}}(\rho)}_{\substack{\text{Ensemble's}\\\text{Expected Loss}}} = \mathbb{E}_\rho[\underbrace{\mathbb{E}_\nu[(y - h_{\boldsymbol{\theta}}(\boldsymbol{x}))^2]}_{\substack{\text{Individual Models'}\\\text{Expected Loss}}}] \quad - \quad \mathbb{E}_\nu[\underbrace{\mathbb{E}_\rho[(h_{\boldsymbol{\theta}}(\boldsymbol{x}) - \mathbb{E}_\rho[h_{\boldsymbol{\theta}}(\boldsymbol{x})])^2]}_{\substack{\text{Variance among}\\\text{individual models}}}]$$

- **Strong ensembles** require strong and diverse individual models: small individual error and high variance.

Measuring Diversity

- For **regression ensembles**, (Krogh and Vedelsby, 1994) showed that:

$$\underbrace{L_{\mathsf{sq}}(\rho)}_{\substack{\text{Ensemble's}\\\text{Expected Loss}}} = \mathbb{E}_\rho[\underbrace{\mathbb{E}_\nu[(y - h_{\boldsymbol{\theta}}(\boldsymbol{x}))^2]}_{\substack{\text{Individual Models'}\\\text{Expected Loss}}}] - \mathbb{E}_\nu[\underbrace{\mathbb{E}_\rho[(h_{\boldsymbol{\theta}}(\boldsymbol{x}) - \mathbb{E}_\rho[h_{\boldsymbol{\theta}}(\boldsymbol{x})])^2]}_{\substack{\text{Variance among}\\\text{individual models}}}]$$

- **Strong ensembles** require strong and diverse individual models: small individual error and high variance.

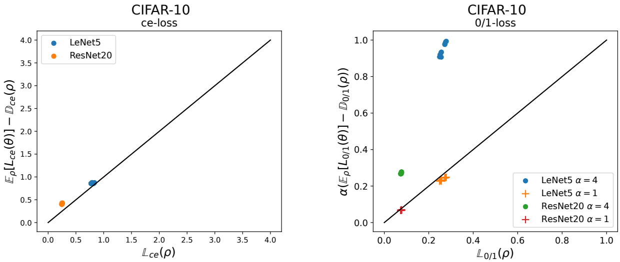
- Existing literature only contains ad-hoc decompositions for other kind of ensembles, but **there is not a general decomposition**.

Empirical validation

Theorem 1

$$\underbrace{L(\rho)}_{\text{Ensemble's Expected Loss}} \leq \alpha \left(\underbrace{\mathbb{E}_\rho[L(\theta)]}_{\text{Individual Models' Expected Loss}} - \underbrace{\mathbb{D}(\rho)}_{\text{Ensemble's Diversity}} \right)$$

where $\alpha = 4$ for the 0/1-loss, otherwise, $\alpha = 1$.



Empirical validation

Theorem 1

$$\underbrace{L(\rho)}_{\substack{\text{Ensemble's}\\ \text{Expected Loss}}} \leq \alpha \left(\underbrace{\mathbb{E}_\rho[L(\theta)]}_{\substack{\text{Individual Models'}\\ \text{Expected Loss}}} - \underbrace{\mathbb{D}(\rho)}_{\substack{\text{Ensemble's}\\ \text{Diversity}}} \right)$$

where $\alpha = 4$ for the $0/1$-loss, otherwise, $\alpha = 1$.