Ensembling Improves the NLL Performance

Ensemble Prediction:

$$P_k^E = \frac{1}{M} \sum_{m=1}^{M} P_{km}, \quad k = 1, \ldots, 10 \quad \text{(for 10 classes)}$$

Negative Log-Likelihood (NLL) Contribution:

$$l_m = -\log\big(P_{km}\big) \quad \text{for model } m$$

$$l^E = -\log\big(P_k^E\big) \quad \text{for the ensemble}$$

Compare Average vs. Ensemble NLL:

$$\bar{l} = \frac{1}{M} \sum_{m=1}^{M} l_m = -\frac{1}{M} \sum_{m=1}^{M} \log\big(P_{km}\big) \geq -\log\Big(\frac{1}{M} \sum_{m=1}^{M} P_{km}\Big) = l^E$$

Reason: Jensen's inequality for the concave function $\log(\cdot)$ implies

$$\log\Big(\tfrac{1}{M} \sum_{m=1}^{M} P_{km}\Big) \;\geq\; \tfrac{1}{M} \sum_{m=1}^{M} \log\big(P_{km}\big).$$