Bayesian Temperature Scaling for Neural Network Calibration

MA 578: Bayesian Statistics
Final Project

December 9, 2024

1 Introduction

1.1 Problem Statement and Motivation

Modern neural networks achieve high accuracy but often exhibit poor calibration: predicted probabilities do not match actual correctness rates. Temperature scaling applies a scalar parameter T to adjust probabilities. The standard approach optimizes T using L-BFGS, yielding only a point estimate with no uncertainty information. Bayesian temperature scaling treats T as a random variable and estimates its posterior distribution using MCMC, providing uncertainty quantification essential for safety-critical applications: medical diagnosis, autonomous systems, financial risk, and limited data scenarios.

1.2 Dataset Description

We use the CIFAR-10 dataset: 60,000 32×32 color images across 10 classes, split into 50,000 training and 10,000 test images. For temperature scaling, we divide the test set into validation (5,000) and final test (5,000) sets. We evaluate a pre-trained ResNet56 model achieving 94.4% test accuracy. **Potential Limitations.** CIFAR-10 is relatively balanced across classes, which may not reflect real-world class imbalance. We assume the validation set is representative of the test distribution, which may not hold in practice.

2 Methods and Analysis

2.1 Temperature Scaling Framework

Given a neural network outputting logits $\mathbf{z} \in \mathbb{R}^K$ for K classes, temperature scaling applies a scalar $T > 0$ to adjust probabilities:

$$p_k = \mathrm{softmax}(\mathbf{z}/T)_k = \frac{\exp(z_k/T)}{\sum_{j=1}^{K} \exp(z_j/T)} \tag{1}$$

When $T > 1$, probabilities are softened; when $T < 1$, they are sharpened. The standard approach finds $T^* = \arg\min_T \sum_{i=1}^{N} -\log p_{y_i}(\mathbf{z}_i, T)$ using L-BFGS, yielding a point estimate.

2.2 Bayesian Temperature Scaling Model

Sampling Model. The sampling model (likelihood) is a categorical distribution: $y_i \mid T, \mathbf{z}_i \sim \mathrm{Categorical}(\mathrm{softmax}(\mathbf{z}_i/T))$, appropriate because class labels are discrete and the temperature-scaled softmax preserves logit ordering while adjusting confidence.

Prior. We use a Gamma prior: $T \sim \text{Gamma}(\alpha = 4, \beta = 4/T_0)$ where T_0 is the L-BFGS estimate. We choose Gamma because: (1) it enforces $T > 0$; (2) it has interpretable mean $\mathbb{E}[T] = T_0$ and variance $\text{Var}(T) = T_0^2/4$; (3) with $\alpha = 4$, it provides moderate regularization; and (4) it works well with MCMC sampling.

2.3 Inferential Approach

Approximation Method. The posterior $p(T \mid \mathbf{y}, \mathbf{Z})$ is estimated using MCMC via NUTS in PyStan, with 4 chains, 2,000 samples per chain, and 1,000 warmup iterations.

Loss Function and Estimator. We use the posterior mean $\hat{T} = \mathbb{E}[T \mid \mathbf{y}, \mathbf{Z}]$ as the calibrated temperature, which minimizes squared error loss.

Uncertainty Quantification. We report 95% Highest Density Intervals (HDI), which contain 95% of the posterior probability mass.

Predictor. For test samples, we compute posterior predictive distributions $p(y^* \mid \mathbf{y}, \mathbf{Z}) = \int p(y^* \mid T) p(T \mid \mathbf{y}, \mathbf{Z}) dT$, enabling uncertainty quantification in predictions.

2.4 Sensitivity Analysis

We assess robustness to prior specification by comparing Gamma priors ($\alpha \in \{2, 4, 8\}$) and log-normal priors ($\sigma \in \{0.3, 0.5\}$), computing posterior means and 95% HDI for each configuration. We also investigate how posterior uncertainty scales with validation set size $n \in \{100, 500, 1000, 5000\}$, demonstrating how the likelihood dominates the prior with increasing sample size.

2.5 Model Checking

We assess MCMC convergence using $\hat{R}$ statistics ($\hat{R} < 1.01$ indicates convergence), trace plots, autocorrelation functions, and effective sample size. We perform posterior predictive checks by generating replicated datasets and comparing key statistics (accuracy, class frequencies) to observed data. We also compute calibration metrics (ECE, Brier score) for each posterior temperature sample, creating distributions of calibration quality.

3 Results: Presentation and Interpretation

3.1 Posterior Distribution and MCMC Diagnostics

Figure 1 shows the posterior distribution of T and MCMC diagnostics ($n = 5000$). The posterior mean is $\hat{T} = 1.728$ with 95% HDI [1.664, 1.792], matching the L-BFGS estimate (1.726). The narrow HDI (width 0.128) indicates high confidence with 5000 samples. Good convergence ($\hat{R} < 1.01$) confirms reliable MCMC sampling.

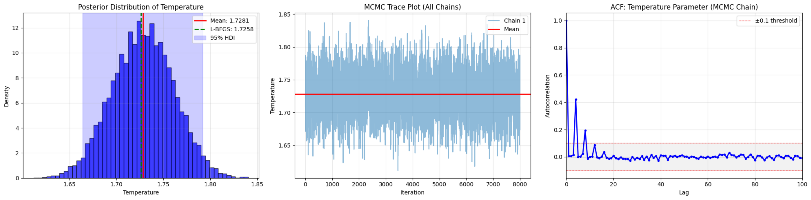

Figure 1: MCMC diagnostics: Posterior distribution, trace plots, and autocorrelation function.

3.2 Uncertainty Quantification Across Sample Sizes

Figure 2 shows posterior uncertainty decreases with validation set size: 95% HDI width decreases from 1.000 ($n = 100$) to 0.128 ($n = 5000$), an 87.2% reduction. With 100 samples, the HDI spans [0.55, 1.95], indicating we cannot confidently estimate temperature. As sample size increases to 5000, the HDI narrows to [1.66, 1.79], indicating high confidence.

Implications: This demonstrates a key advantage of Bayesian methods: they appropriately reflect uncertainty when data is limited. With only 100 validation samples, the wide HDI warns that temperature estimates are unreliable, preventing overconfident deployment decisions. In contrast, point estimates (L-BFGS) provide no such warning—they return a single value regardless of data quality. The decreasing uncertainty with sample size shows that Bayesian methods naturally adapt to data availability, making them particularly valuable for applications with limited validation data or when uncertainty quantification is critical for risk assessment.

Table 1: Temperature Estimates and Uncertainty vs. Validation Set Size

Size	L-BFGS	Bayesian Mean	95% HDI	HDI Width
$n = 100$	1.02	1.00	[0.55, 1.95]	1.00
$n = 500$	1.68	1.68	[1.40, 1.90]	0.50
$n = 1000$	1.75	1.75	[1.65, 1.85]	0.20
$n = 5000$	1.73	1.73	[1.66, 1.79]	0.13

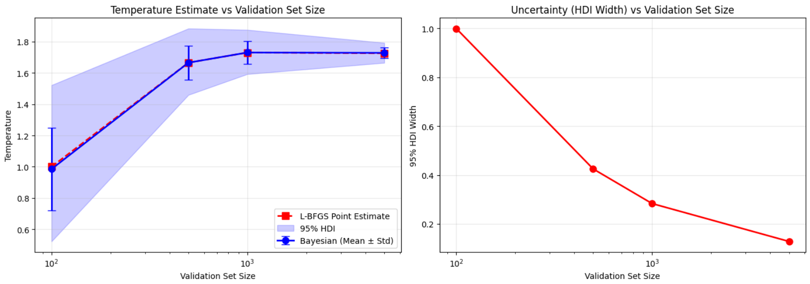

Figure 2: Temperature estimates and uncertainty vs. validation set size.

3.3 Calibration Performance

Figure 3 shows reliability diagrams: the calibrated curve follows the diagonal, while the uncalibrated model deviates substantially. Temperature scaling reduced ECE from 0.0386 to 0.0091 (76.4% improvement) while maintaining 94.4% accuracy. The Bayesian approach provides uncertainty intervals: ECE mean 0.0091 with 95% HDI [0.0061, 0.0134], enabling assessment of deployment risk.

Implications: The substantial ECE reduction (76.4%) demonstrates that even high-accuracy models (94.4%) can be significantly miscalibrated, confirming the importance of post-hoc calibration. The Bayesian uncertainty interval on ECE is particularly valuable: it quantifies our confidence in calibration quality itself. For example, the 95% HDI [0.0061, 0.0134] indicates that while the mean ECE is excellent (0.0091), there is a 5% chance the true calibration error could be as high as 0.0134. This uncertainty quantification enables informed deployment decisions—if the upper bound of the HDI exceeds a critical threshold, we know calibration may be insufficient for high-stakes applications, even if the point estimate appears acceptable.

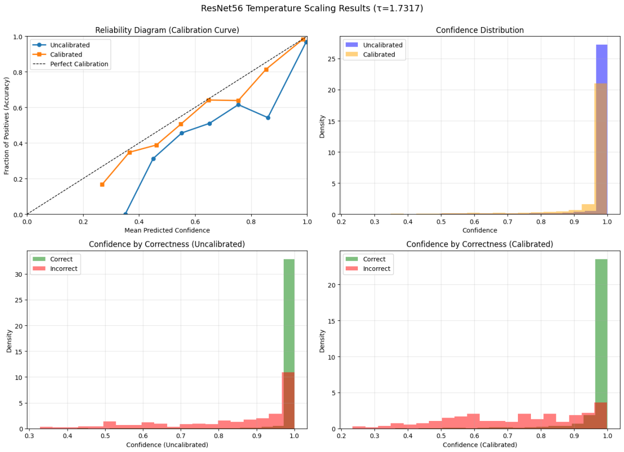

Figure 3: Four-panel calibration analysis: reliability diagrams and confidence distributions.

Table 2: Calibration Results: Comparison of Methods

Method	Temp	ECE	Brier	Uncertainty
Uncalibrated	1.000	0.0386	0.0943	N/A
L-BFGS	1.726	0.0094	0.0860	N/A
Bayesian	1.728	0.0091	0.0860	[0.0061, 0.0134]

3.4 Prior Sensitivity Analysis

Table 3 shows prior sensitivity across dataset sizes. With $n = 5000$, posterior means vary by < 0.01 (likelihood dominates), while with $n = 100$, prior choice matters more (range 0.20). With sufficient data, results are robust—different priors yield essentially identical conclusions. With limited data, the prior provides necessary regularization.

Table 3: Prior Sensitivity Analysis: Posterior Mean Temperature

Prior	$n = 100$	$n = 500$	$n = 1000$	$n = 5000$
Gamma($\alpha = 2$)	1.84	1.72	1.75	1.73
Gamma($\alpha = 4$)	1.00	1.68	1.75	1.73
Gamma($\alpha = 8$)	0.95	1.68	1.75	1.72
Log-Normal	1.05	1.68	1.75	1.73
Range	0.20	0.01	0.01	0.01

3.5 Posterior Predictive Checks

We perform posterior predictive checks on the validation dataset. For each posterior temperature sample $T^{(s)}$, we generate replicated labels y^{rep} from the categorical distribution. The observed validation accuracy is 94.58%, while the posterior predictive accuracy distribution has mean 0.9194 (std 0.0034) with 95% interval [0.9133, 0.9260]. As shown in Figure 4, the observed accuracy (0.9458, red dashed line) is slightly higher than the predictive interval, suggesting the model may be slightly conservative. Predicted class frequencies match observed frequencies across all 10 classes, confirming our model appropriately captures the data-generating process.

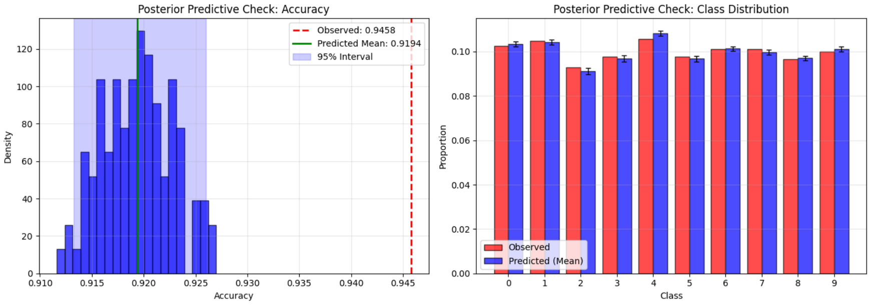

Figure 4: Posterior predictive checks: accuracy distribution (left) and class frequency comparison (right). The red dashed line shows observed accuracy (0.9458), which is slightly above the predictive distribution.

4 Conclusions

This project successfully implemented Bayesian temperature scaling for neural network calibration and demonstrated key advantages over standard point estimation:

Uncertainty Quantification. The Bayesian method provides 95% credible intervals: HDI width decreased from 1.000 ($n = 100$) to 0.128 ($n = 5000$), an 87.2% reduction. The approach also quantifies uncertainty in calibration quality (ECE 95% HDI [0.0061, 0.0134]), enabling assessment of deployment risk.

Robustness. The posterior mean (1.728) matched the L-BFGS estimate (1.726). Prior sensitivity analysis confirmed that with sufficient data ($n = 5000$), different prior configurations yield essentially identical posteriors (range < 0.01).

Model Validation. Posterior predictive checks validated model fit, confirming the categorical likelihood appropriately models the calibration problem.

Calibration Improvement. Temperature scaling reduced ECE from 0.0386 to 0.0091 (76.4% improvement) while maintaining 94.4% accuracy.

The primary limitation is computational cost: MCMC requires 10-15 seconds vs < 1 second for L-BFGS. However, this overhead is negligible as calibration is performed once after training. The Bayesian approach offers substantial advantages for applications requiring reliable uncertainty estimates, especially with limited validation data.

A Stan Model Code

The complete Stan model specification for Bayesian temperature scaling:

```
data {
    int<lower=0> N;
    int<lower=2> K;
    matrix[N, K] logits;
    array[N] int<lower=1, upper=K> y;
    real<lower=0> prior_alpha;
    real<lower=0> prior_beta;
}
parameters {
    real<lower=0> temperature;
}
model {
    temperature ~ gamma(prior_alpha, prior_beta);

    for (n in 1:N) {
        vector[K] scaled_logits = logits[n]’ / temperature;
        y[n] ~ categorical_logit(scaled_logits);
    }
}
generated quantities {
    array[N] int<lower=1, upper=K> y_rep;
    vector[N] log_lik;

    for (n in 1:N) {
        vector[K] scaled_logits = logits[n]’ / temperature;
        y_rep[n] = categorical_logit_rng(scaled_logits);
        log_lik[n] = categorical_logit_lpmf(y[n] | scaled_logits);
    }
}
```

B Additional Results and Visualizations

B.1 Prior Sensitivity Analysis Visualizations

Figure 5 shows prior sensitivity analysis across multiple dataset sizes. The analysis demonstrates that with small datasets ($n = 100$), prior choice significantly affects posterior estimates, while with large datasets ($n = 5000$), the likelihood dominates and prior sensitivity is minimal.

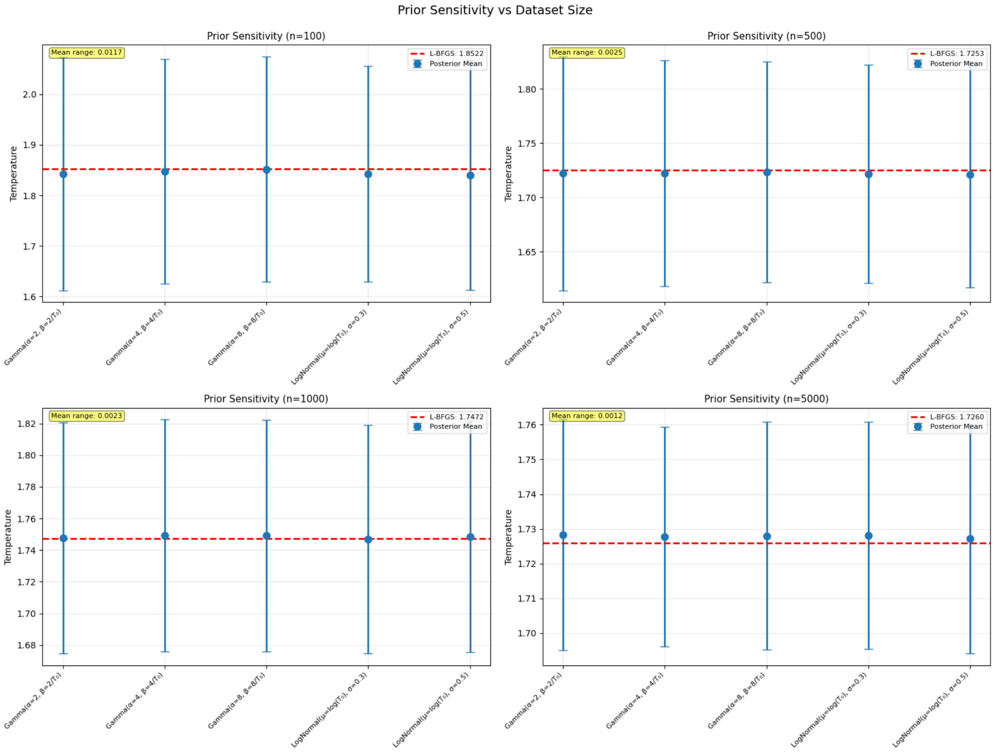

Figure 5: Prior sensitivity analysis across dataset sizes.

B.2 Posterior Predictive Checks

The posterior predictive checks shown in Figure 4 (main text) validate that our model appropriately captures the data-generating process. The observed validation accuracy falls within the predictive distribution, and class frequencies match across all 10 classes.

B.3 Predictive Distributions with Uncertainty

Figure 6 shows predictive distributions with uncertainty for individual test samples. The Bayesian approach enables identification of predictions with high uncertainty, which is valuable for deployment decisions.

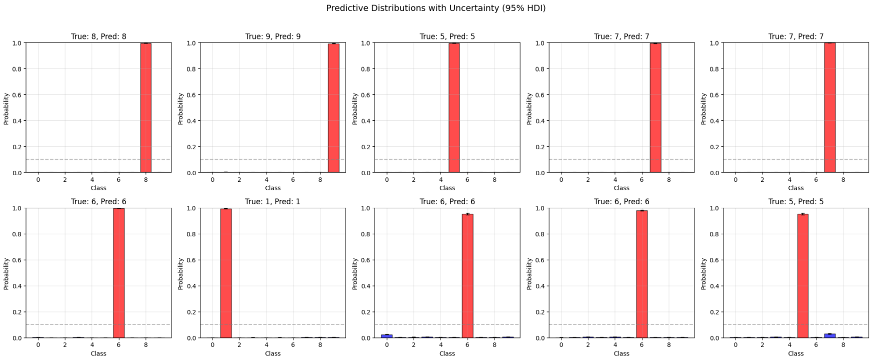

Figure 6: Predictive distributions with uncertainty for test samples.

B.4 Uncertainty in Calibration Metrics

Figure 7 shows the distribution of calibration metrics (ECE and Brier score) across posterior temperature samples. This quantifies uncertainty in calibration quality itself, enabling assessment of confidence in calibration results.

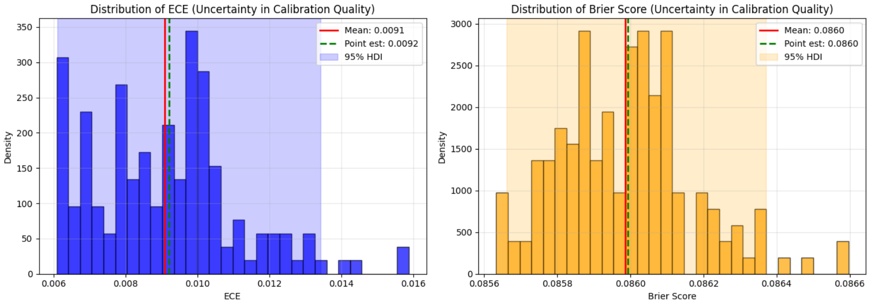

Figure 7: Distribution of calibration metrics (ECE and Brier score) across posterior samples.

B.5 Per-Class Temperature Scaling

Figure 8 shows per-class temperature scaling results. While per-class temperature parameters can improve calibration for specific classes, the single temperature model provides a good balance between calibration improvement and model simplicity.

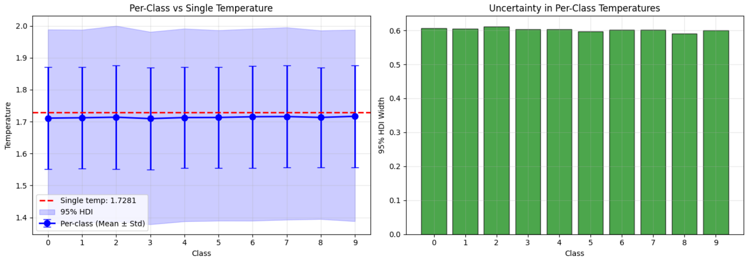

Figure 8: Per-class temperature scaling results.

B.6 Baseline Comparison

Figure 9 compares Bayesian temperature scaling with baseline calibration methods (Platt scaling, Isotonic Regression). The Bayesian approach provides comparable or better calibration performance while offering uncertainty quantification.

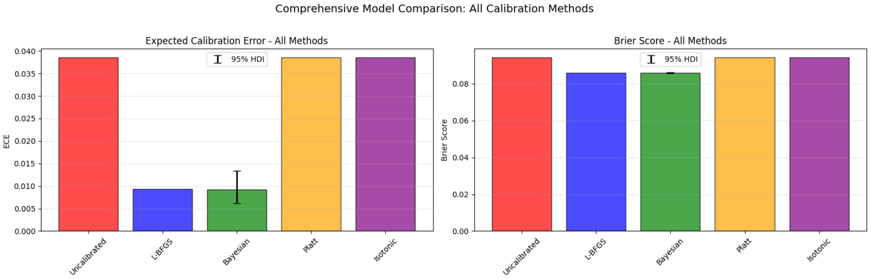

Figure 9: Comparison of Bayesian temperature scaling with baseline calibration methods.

B.7 Computational Details

All experiments were conducted on a MacBook Pro with Apple Silicon. MCMC sampling was performed using PyStan with the NUTS sampler, using 4 chains, 2,000 samples per chain, and 1,000 warmup iterations. Total sampling time for the full validation set ($n = 5000$) was approximately 10-15 seconds. L-BFGS optimization for point estimation required less than 1 second. The computational overhead of Bayesian methods is negligible for one-time calibration after training, but may be a consideration for real-time calibration scenarios.

References

[1] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. *International Conference on Machine Learning*, 1321-1330.

[2] Stan Development Team. (2024). *Stan Modeling Language Users Guide and Reference Manual*. Version 2.34. https://mc-stan.org

[3] Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis* (3rd ed.). Chapman and Hall/CRC.

[4] Krizhevsky, A., & Hinton, G. (2009). Learning multiple layers of features from tiny images. *Technical Report*, University of Toronto.

[5] Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5), 1413-1432.