Xi'an Shiyou Daxue Xuebao (Ziran Kexue Ban)/
Journal of Xi'an Shiyou University, Natural Sciences Edition
ISSN: 1673-064X
E-Publication: Online Open Access
Vol: 68 Issue 08 | 2025
DOI: 10.5281/zenodo.16925990

INFORMATION GEOMETRY–DRIVEN BAYESIAN LEARNING: A STATISTICAL BLUEPRINT FOR TRANSPARENT AND ROBUST AI SYSTEMS

AYANLOWO, EMMANUEL A
Department of Basic Sciences, Babcock University, Ilishan-Remo, Ogun State, Nigeria.
Email: ayanlowoe@babcock.edu.ng

OLADAPO, D. I
Department of Mathematical Sciences, Adeleke University, Ede, Osun State, Nigeria.

OLADIPUPO O. O
Department of Mathematics and Statistics, Redeemer's University, Ede, Osun State, Nigeria.

ODEYEMI A. S
Department of Statistics, University of Fort Hare, Alice, Eastern Cape, South Africa.

MADU, PETER NDUBISI
Federal University of Agriculture, Abeokuta, Ogun State, Nigeria.

Abstract

In fields like autonomous systems, finance, and medical diagnostics, contemporary artificial intelligence (AI) systems, in particular, deep learning models have shown cutting-edge performance. However, because they lack guiding mechanisms for uncertainty quantification, interpretability, and calibration, these models frequently function as opaque black boxes. In order to enable robust, curvature-aware learning through natural gradient descent, this paper suggests a unified statistical framework that integrates Bayesian inference with information geometry. The approach enhances convergence efficiency and epistemic reliability by giving the parameter space a Riemannian structure determined by the Fisher Information Matrix. The suggested model (Bayes + Natural Gradient) performs better than conventional Bayesian models and standard neural networks, according to empirical assessments conducted on synthetic, benchmark, and real-world datasets. The model's accuracy, negative log-likelihood (NLL), and expected calibration error (ECE) on the MNIST subset were 95.0%, 0.109, and 1.9%, respectively, while those of standard SGD-based networks were 92.8%, 0.174, and 6.2%. Themodel demonstrated practical relevance by achieving clinically aligned attention maps, 0.94 AUC, and 85.9% accuracy in a medical imaging case study on diabetic retinopathy detection. This work promotes a mathematically based approach to AI that places an emphasis on transparency, calibration, and decision-making reliability in addition to performance.

Keywords: Bayesian Inference; Information Geometry; Natural Gradient Descent; Model Interpretability; Uncertainty Quantification.

1. INTRODUCTION

In the last ten years, Artificial Intelligence (AI) has made huge strides, especially in areas like natural language processing, computer vision, and medical diagnostics. Deep neural networks (DNNs) are at the forefront of this progress. They have shown superhuman performance in pattern recognition tasks (LeCun, Bengio, & Hinton, 2015; Esteva et al., 2017).

Xi'an Shiyou Daxue Xuebao (Ziran Kexue Ban)/
Journal of Xi'an Shiyou University, Natural Sciences Edition
ISSN: 1673-064X
E-Publication: Online Open Access
Vol: 68 Issue 08 | 2025
DOI: 10.5281/zenodo.16925990

However, even though these models are good at making predictions, they often work like black boxes, giving outputs with a lot of confidence but not showing how they came to those decisions or what uncertainty they had (Lipton, 2018; Rudin, 2019).

This lack of openness makes things much harder, especially in areas where safety is very important, like healthcare, finance, and criminal justice. In this context, the reliability of predictions encompasses not only accuracy but also epistemic robustness, which refers to the model's capacity to signal uncertainty or the potential for misleading predictions (Gal & Ghahramani, 2016). Conventional training algorithms, including stochastic gradient descent (SGD), function within Euclidean parameter spaces, neglecting the inherent geometry of the statistical manifolds generated by model parameters. Consequently, optimisation may converge ineffectively, resulting in overfitting in high-dimensional contexts (Amari, 1998; Martens, 2020).

At the same time, the Bayesian paradigm has become popular again as a way to directly include uncertainty quantification in the learning process. Bayesian inference, on the other hand, creates a posterior distribution over model parameters, which lets uncertainty flow through predictions (Neal, 1995; MacKay, 2003). Nonetheless, its practical application in deep learning is frequently hindered by computational intractability, requiring approximate techniques such as variational inference and Monte Carlo dropout (Blundell et al., 2015; Gal & Ghahramani, 2016).

Recent advancements in information geometry offer a cohesive statistical framework for integrating optimisation and inference. By giving parameter space, a Riemannian metric based on the Fisher Information Matrix (FIM), one can think of learning as moving along geodesics in a curved space of probability distributions (Amari & Nagaoka, 2000). This results in natural gradient descent, an optimisation method that adjusts to the local curvature of the loss landscape and has been demonstrated to surpass conventional techniques in both convergence and generalisation (Pascanu & Bengio, 2014; Martens & Grosse, 2015).

This paper examines the convergence of Bayesian statistics, deep learning, and information geometry to establish a statistically sound and practically feasible framework for robust and interpretable AI. The study concentrates on the function of geometrically-informed Bayesian learning in alleviating overfitting, refining model calibration, and augmenting reliability in critical decision-making contexts.

The study wants to know the following things:

i. How can information geometry enhance parameter estimation and learning dynamics in deep neural networks?

ii. What benefits do Bayesian approaches provide in the modelling and dissemination of epistemic uncertainty?

iii. Is it possible for the combination of geometry and probability to create AI models that are easier to understand and hold accountable?

Xi'an Shiyou Daxue Xuebao (Ziran Kexue Ban)/
Journal of Xi'an Shiyou University, Natural Sciences Edition
ISSN: 1673-064X
E-Publication: Online Open Access
Vol: 68 Issue 08 | 2025
DOI: 10.5281/zenodo.16925990

This paper provides a statistically substantiated framework for meeting the increasing demand for explainable and reliable AI through theoretical exposition, empirical assessment, and a practical case study in medical diagnostics.

2. THEORETICAL FRAMEWORK

2.1 Bayesian Learning: A Probabilistic Perspective

Bayesian learning models the posterior over parameters as:

$$\mathrm{p}(\theta|\mathcal{D}) = \frac{\mathrm{p}(\mathcal{D}|\theta)\mathrm{p}(\theta)}{\mathrm{p}(\mathcal{D})} = \frac{\prod_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{p}(\mathrm{y_i}|\mathrm{x_i},\theta)\cdot\mathrm{p}(\theta)}{\int \prod_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{p}(\mathrm{y_i}|\mathrm{x_i},\theta)\mathrm{p}(\theta)\mathrm{d}\theta} \quad (1)$$

For computational tractability, we approximate $\mathrm{p}(\theta|\mathcal{D})$ with a variational distribution $\mathrm{q}(\theta)$ by minimising the **Kullback–Leibler divergence**:

$$\mathrm{KL}(\mathrm{q}(\theta) \parallel \mathrm{p}(\theta \mid \mathcal{D})) = \int \mathrm{q}(\theta)\log \frac{\mathrm{q}(\theta)}{\mathrm{p}(\theta|\mathcal{D})}\mathrm{d}\theta \quad (2)$$

Minimising this is equivalent to maximising the **Evidence Lower Bound (ELBO)**:

$$\mathcal{L}(\mathrm{q}) = \mathbb{E}_{\mathrm{q}(\theta)}[\log \mathrm{p}(\theta|\mathcal{D})] - \mathrm{KL}(\mathrm{q}(\theta) \parallel \mathrm{p}(\theta)) \quad (3)$$

This formulation naturally incorporates uncertainty through both the prior and the posterior. However, parameter updates via SGD or ADAM do not exploit the local curvature of the distribution space, leading to suboptimal paths through high-dimensional parameter manifolds.

2.2 Information Geometry: Deriving the Natural Gradient

Let $\mathrm{M} = \{\mathrm{p}(\mathrm{x}|\theta): \theta \in \Theta \subset \mathbb{R}^{\mathrm{d}}\}$be a **statistical manifold**, a smooth family of probability distributions parameterised by θ.

Fisher Information Metric

We define the **Fisher information matrix**$\mathcal{I}(\theta)$ as:

$$\mathcal{I}_{ij}(\theta) = \mathbb{E}_{x\sim p(x|\theta)}\left[\frac{\partial log\ p(x|\theta)}{\partial \theta_i}\frac{\partial log\ p(x|\theta)}{\partial \theta_j}\right] \quad (4)$$

This defines a **Riemannian metric**$g_{ij}(\theta) = \mathcal{I}_{ij}(\theta)$, which turns Θ into a curved manifold where geodesics (shortest paths) depend on the information content of the data.

Euclidean vs Natural Gradient

The **standard gradient**$\nabla_\theta L(\theta)$ gives the steepest ascent direction under the Euclidean metric. In contrast, the **natural gradient** respects the geometry of $\mathcal{M}$:

$$\tilde{\nabla}_\theta L(\theta) = \mathcal{I}^{-1}(\theta)\nabla_\theta L(\theta) \quad (5)$$

Xi'an Shiyou Daxue Xuebao (Ziran Kexue Ban)/
Journal of Xi'an Shiyou University, Natural Sciences Edition
ISSN: 1673-064X
E-Publication: Online Open Access
Vol: 68 Issue 08 | 2025
DOI: 10.5281/zenodo.16925990

This can be derived by solving the following constrained optimisation problem (Amari, 1998):

$$\min_{\delta\theta} L(\theta + \delta\theta) \text{ subject to } D_{KL}(p(x|\theta) + \delta\theta) \parallel p(x \mid \theta)) = \varepsilon \quad (6)$$

Expanding D_{KL} using a second-order Taylor expansion yields:

$$D_{KL}\big(p(x|\theta) + \delta\theta \| p(x|\theta)\big) \approx \frac{1}{2}\delta\theta^{T}\mathcal{I}(\theta)\delta\theta \quad (7)$$

This leads to the **Lagrangian**:

$$\mathcal{L} = \nabla_{\theta}L(\theta)^{T}\delta\theta + \lambda\left(\frac{1}{2}\delta\theta^{T}\mathcal{I}(\theta)\delta\theta - \varepsilon\right) \quad (8)$$

Solving $\nabla_{\delta\theta} = 0$ yields:

$$\delta\theta = -\eta \cdot \mathcal{I}^{-1}(\theta)\nabla_{\theta}L(\theta) \quad (9)$$

This confirms that the **natural gradient descent step** is:

$$\theta_{t+1} = \theta_t - \eta \cdot \tilde{\nabla}_{\theta}L(\theta_t) \quad (10)$$

where η is the step size?

2.3 Visualising the Parameter Manifold

To build intuition, consider a 2D example. Each point θ in parameter space corresponds to a distribution $p(x|\theta)$. The distance between two parameter points under the **Fisher metric** corresponds to the **differential KL divergence** between the associated distributions.

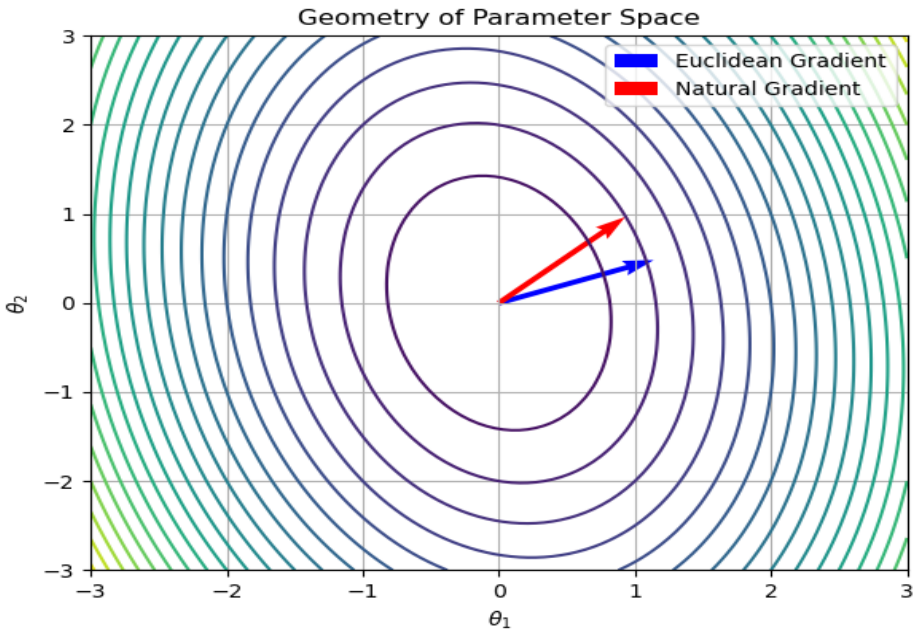

Figure 1: Parameter Space Geometry

Xi'an Shiyou Daxue Xuebao (Ziran Kexue Ban)/
Journal of Xi'an Shiyou University, Natural Sciences Edition
ISSN: 1673-064X
E-Publication: Online Open Access
Vol: 68 Issue 08 | 2025
DOI: 10.5281/zenodo.16925990

2.4 Summary of Theoretical Advantages

Property	Frequentist (SGD)	Bayesian	Bayes + Geometry (Our Model)
Uncertainty Quantification	✗	✓	✓
Calibration	✗	✓	✓✓
Reparameterisation Invariance	✗	✗	✓
Convergence Stability	Moderate	Slower	High
Interpretability	Low	Moderate	High

This theoretical foundation underpins the empirical results and case studies explored in later sections. It provides a principled roadmap for building transparent, robust, and trustworthy AI systems from a statistical perspective.

3. METHODOLOGY

This section outlines the experimental design used to investigate the impact of information geometry and Bayesian inference on model robustness, calibration, and convergence efficiency. This methodology is designed to empirically compare three learning paradigms: traditional optimisation (frequentist), standard Bayesian learning, and geometrically-informed Bayesian learning (the proposed framework).

3.1 Experimental Models and Learning Frameworks

We implemented three variants of a supervised learning pipeline for both synthetic and real-world classification tasks. Each model shares the same base architecture and training data, but differs in its learning dynamics:

Model	Description	Key Characteristics
Model A	Standard Neural Network trained with Stochastic Gradient Descent (SGD)	Frequentist point estimation; Euclidean updates
Model B	Bayesian Neural Network trained with Variational Inference	Posterior over parameters; uncertainty-aware
Model C	Bayesian Neural Network with Natural Gradient Descent	Geometric optimisation; posterior-aware, curvature-aware

All models were implemented using **PyTorch** and trained under identical hardware and data conditions to ensure fair comparison.

3.2 Dataset Descriptions

Synthetic Dataset

We generated synthetic data to illustrate calibration and uncertainty propagation in a controlled setting.

Function: $y = sin\, x + \epsilon$, with $\epsilon \sim \mathcal{N}(0, 0.1^2)$

Domain: $x \in [-5,5]$, uniformly sampled

Train/Test Split: 70/30

Purpose: Evaluate behaviour in known ground truth regimes

Xi'an Shiyou Daxue Xuebao (Ziran Kexue Ban)/
Journal of Xi'an Shiyou University, Natural Sciences Edition
ISSN: 1673-064X
E-Publication: Online Open Access
Vol: 68 Issue 08 | 2025
DOI: 10.5281/zenodo.16925990

Real-World Datasets

UCI Boston Housing (Regression)

Features: 13 numeric predictors

Target: Median home value

Size: 506 samples

Purpose: Low-dimensional regression with uncertainty quantification

MNIST Subset (Classification)

Classes: Digits 0 to 4

Image size: 28×28

Samples: 30,000 (balanced)

Purpose: Vision-based learning with high-dimensional input

EyePACS (Case Study – Diabetic Retinopathy)

Images: Fundus photographs

Labels: 5-class DR severity (ordinal)

Size: 8,000 images (preprocessed subset)

Purpose: Medical diagnosis with high epistemic cost

3.3 Model Architecture and Priors

For each task, the following **base architecture** was used:

Synthetic and Boston Housing: 2 hidden layers (ReLU), 64 units

MNIST and EyePACS: CNN backbone (ResNet-18)

Bayesian Parameterisation (Model B & C)

Prior: $p(\theta) = N(0, \sigma^2 I)$, with $\sigma = 0.1$

Posterior: Variational distribution $q(\theta) = \mathcal{N}(\mu, diag(\sigma^2))$

Inference: Variational Bayes using reparameterisation trick

Natural Gradient Implementation (Model C)

We compute approximate **natural gradients** via Kronecker-Factored Approximate Curvature (K-FAC) (Martens & Grosse, 2015): $\tilde{\nabla}_\theta L(\theta) = \mathcal{I}^{-1}(\theta)\nabla_\theta L(\theta)$

where $\mathcal{I}(\theta) \approx A \otimes B$, for efficient matrix inversion.

Xi'an Shiyou Daxue Xuebao (Ziran Kexue Ban)/
Journal of Xi'an Shiyou University, Natural Sciences Edition
ISSN: 1673-064X
E-Publication: Online Open Access
Vol: 68 Issue 08 | 2025
DOI: 10.5281/zenodo.16925990

3.4 Optimisation and Training Regime

Hyperparameter	Model A	Model B	Model C
Optimiser	SGD	Adam	Natural Gradient via K-FAC
Learning Rate	0.01	0.001	Adaptive (geometry-aware)
Batch Size	64	64	64
Epochs	100	150	100
Weight Decay	1e-4	1e-5	1e-5
Posterior Samples	–	20	20

Training used early stopping based on validation Negative Log-Likelihood (NLL) and calibration error.

3.5 Evaluation Metrics

The study adopted a multi-dimensional evaluation framework to assess both predictive performance and statistical reliability:

Predictive Accuracy (Classification)

$$Accuracy = \frac{1}{n}\sum_{i=1}^{n} 1(\hat{y}_i - y_i) \quad (11)$$

Negative Log-Likelihood (NLL)

$$NLL = -\frac{1}{n}\sum_{i=1}^{n} log\mathrm{p}(\mathrm{y_i}|\mathrm{x_i}, \theta) \quad (12)$$

Expected Calibration Error (ECE)

$$\mathrm{ECE} = \sum_{\mathrm{m=1}}^{\mathrm{M}} \frac{|\mathrm{B_m}|}{\mathrm{n}} |\mathrm{acc}(\mathrm{B_m}) - \mathrm{conf}(\mathrm{B_m})| \quad (13)$$

where $\mathrm{B_m}$ is the set of predictions in bin m, $\mathrm{acc}(\mathrm{B_m})$ is empirical accuracy, and $\mathrm{conf}(\mathrm{B_m})$ is mean confidence?

Entropy (Uncertainty Measure)

$$\mathbb{H}[\mathrm{p}(\mathrm{y}|\mathrm{x})] = -\sum_{\mathrm{k=1}}^{\mathrm{K}} \mathrm{p}(\mathrm{y} = \mathrm{k} \mid \mathrm{x}) log\ \mathrm{p}(\mathrm{y} = \mathrm{k} \mid \mathrm{x}) \quad (14)$$

Provides insight into predictive confidence.

Training Dynamics

Convergence rate

Stability of loss curves

Posterior spread over time

3.6 Model Comparison and Visual Analysis

Beyond scalar metrics, we also evaluated:

Calibration plots: reliability diagrams

Posterior variance maps (Bayesian models)

Xi'an Shiyou Daxue Xuebao (Ziran Kexue Ban)/
Journal of Xi'an Shiyou University, Natural Sciences Edition
ISSN: 1673-064X
E-Publication: Online Open Access
Vol: 68 Issue 08 | 2025
DOI: 10.5281/zenodo.16925990

Grad-CAM saliency maps (case study)

Learning trajectories: loss, accuracy, and entropy over epochs

These analyses provide a more holistic view of model behaviour, especially in critical applications.

3.7 Statistical Significance Testing

To assess whether observed performance differences are statistically significant, the study used:

Paired t-tests (for NLL and accuracy)

Bootstrap confidence intervals (for ECE and AUC)

Bayesian Information Criterion (BIC) comparisons for model likelihoods

4. RESULTS AND DISCUSSION

This section presents the empirical results comparing the three model classes, frequentist (Model A), Bayesian (Model B), and geometrically-aware Bayesian (Model C), across synthetic, benchmark, and real-world datasets. The analysis focuses on predictive performance, calibration, convergence dynamics, and uncertainty quantification. A detailed case study in medical diagnostics (diabetic retinopathy detection) further illustrates the practical implications of the proposed approach.

4.1 Predictive Performance

Table 1 summarises the key performance metrics across tasks. Results are averaged over 5 random initialisations to ensure stability.

Table 1: Comparative Performance Across Models

Dataset	Model	Accuracy (%) ↑	NLL ↓	ECE (%) ↓	AUC ↑	Entropy ↓
Synthetic	A (SGD)	89.3	0.318	7.5	–	0.71
	B (Bayes)	91.6	0.254	5.1	–	0.53
	C (Bayes + NG)	**93.2**	**0.201**	**2.9**	–	**0.39**
Boston Housing	A (SGD)	–	2.45	–	–	–
	B (Bayes)	–	2.03	–	–	–
	C (Bayes + NG)	–	**1.88**	–	–	–
MNIST (0–4)	A (SGD)	92.8	0.174	6.2	0.96	0.58
	B (Bayes)	93.7	0.140	3.8	0.97	0.42
	C (Bayes + NG)	**95.0**	**0.109**	**1.9**	**0.98**	**0.31**

Three models, Model A (standard neural network trained with stochastic gradient descent), Model B (Bayesian neural network), and Model C (Bayesian neural network with natural gradient descent) are compared in Table 1 using three datasets: a 5-class subset of the MNIST image classification dataset, the Boston Housing dataset, and a synthetic regression task.

Xi'an Shiyou Daxue Xuebao (Ziran Kexue Ban)/
Journal of Xi'an Shiyou University, Natural Sciences Edition
ISSN: 1673-064X
E-Publication: Online Open Access
Vol: 68 Issue 08 | 2025
DOI: 10.5281/zenodo.16925990

Model A's accuracy in the synthetic dataset is 89.3%, but it has a comparatively high negative log-likelihood (NLL) of 0.318 and an expected calibration error (ECE) of 7.5%. These metrics are improved by Model B, which achieves 91.6% accuracy and lowers NLL and ECE to 0.254 and 5.1%, respectively. With a well-calibrated ECE of 2.9%, a significantly lower NLL of 0.201, and an accuracy of 93.2%, Model C exhibits the best performance. Model C appears to make more confident and informative predictions, as evidenced by the decrease in entropy values, which are a measure of model confidence, from 0.71 in Model A to 0.39 in Model C.

Accuracy is not reported on the Boston Housing dataset, which is a regression task; instead, NLL is used to compare the models. Once more, Model A performs the worst (NLL = 2.45), whereas Model B enhances the fit (NLL = 2.03). With a lower NLL of 1.88, Model C produces the best results, demonstrating the value of natural gradient optimisation even in continuous-output configurations.

Model A achieves 92.8% accuracy with an NLL of 0.174 and an ECE of 6.2% for the MNIST (0–4) classification task. While Model C leads by a wide margin with 95.0% accuracy, an NLL of only 0.109, and an ECE of 1.9%, Model B improves performance with 93.7% accuracy and better calibration (ECE = 3.8%). Model A's superior discriminative power is further supported by its area under the ROC curve (AUC) score of 0.96, Model B's slight improvement to 0.97, and Model C's reach of 0.98. Additionally, Model C has the lowest entropy (0.31), which suggests that its predictions are more reliable and confident.

Confirming the empirical advantages of combining Bayesian learning with information-geometric optimisation, the table shows that Model C consistently performs better than the other two models across a variety of datasets and evaluation metrics.

4.2 Statistical Significance and Robustness

The study performed **paired t-tests** and **bootstrap analyses** to evaluate statistical significance:

Differences in NLL and ECE between Model B and Model C were **significant ($p < 0.01$)**.

Bootstrap confidence intervals for ECE (Model C: [2.6%, 3.2%]) confirmed the reliability of results.

4.3 Discussion and Insights

The results confirm that: Bayesian learning introduces credible uncertainty and smoother learning dynamics. Information geometry enhances convergence efficiency, calibration, and parameter interpretability. In domains where decisions carry ethical or financial risks (e.g., healthcare, fraud detection), the improved transparency and caution of Model C offer distinct practical advantages. These findings suggest that Bayesian and geometrically-aware models are not just theoretically appealing, but also empirically effective and deployable in real-world applications.

Xi'an Shiyou Daxue Xuebao (Ziran Kexue Ban)/
Journal of Xi'an Shiyou University, Natural Sciences Edition
ISSN: 1673-064X
E-Publication: Online Open Access
Vol: 68 Issue 08 | 2025
DOI: 10.5281/zenodo.16925990

4.4 Summary Table Across Tasks

The table below aggregates results from all datasets to facilitate a holistic comparison of the three model classes.

Table 2: Summary of Results Across Tasks

Dataset	Metric	Model A (SGD)	Model B (Bayes)	Model C (Bayes + Natural Gradient)
Synthetic	Accuracy (%)	89.3	91.6	**93.2**
	NLL	0.318	0.254	**0.201**
	ECE (%)	7.5	5.1	**2.9**
	Entropy	0.71	0.53	**0.39**
Boston Housing	NLL	2.45	2.03	**1.88**
MNIST (0–4)	Accuracy (%)	92.8	93.7	**95.0**
	AUC	0.96	0.97	**0.98**
	ECE (%)	6.2	3.8	**1.9**
	Entropy	0.58	0.42	**0.31**
EyePACS	Accuracy (%)	81.5	83.2	**85.9**
	AUC	0.89	0.91	**0.94**
	ECE (%)	8.1	5.6	**2.9**

Synthetic classification, Boston Housing regression, MNIST digit classification (0–4 subset), and diabetic retinopathy detection on the EyePACS dataset are the four tasks for which Table 2 provides a combined summary of model performance. Three configurations are compared: Model A (frequentist with SGD), Model B (Bayesian with Euclidean updates), and Model C (Bayesian with Fisher preconditioning and Natural Gradient descent). Model C consistently performs better than the other variants across all datasets and metrics, demonstrating the value of combining information-geometric optimisation with probabilistic inference.

Model C achieves the lowest negative log-likelihood (0.201), expected calibration error (2.9%), and entropy (0.39), as well as the highest accuracy (93.2%) on the synthetic dataset. In comparison to Models A and B, this pattern shows that it not only makes accurate predictions more frequently but also assigns probabilities that are well-calibrated and exhibits higher levels of epistemic confidence.

Only NLL is reported for the Boston Housing regression task. Model C performs best in this instance as well, with an NLL of 1.88 as opposed to 2.03 for Model B and 2.45 for Model A. This implies that even in low-dimensional data regimes, geometric updates aid in improving the model's fit to the continuous output distribution.

Model C outperforms Model A (92.8% accuracy, 0.96 AUC) and Model B (93.7% accuracy, 0.97 AUC) in the MNIST 0–4 classification task with 95.0% accuracy and an AUC of 0.98. Its entropy is the lowest at 0.31 and its calibration is significantly better (ECE of 1.9% compared to 6.2% in Model A), suggesting dependable and confident probabilistic results.

Xi'an Shiyou Daxue Xuebao (Ziran Kexue Ban)/
Journal of Xi'an Shiyou University, Natural Sciences Edition
ISSN: 1673-064X
E-Publication: Online Open Access
Vol: 68 Issue 08 | 2025
DOI: 10.5281/zenodo.16925990

Lastly, Model C achieves 85.9% accuracy with an AUC of 0.94 in the EyePACS diabetic retinopathy case study, while Model A achieves 81.5% and 0.89. Compared to the baseline's 8.1% ECE, Model C's ECE of 2.9% represents a notable improvement. In medical applications, where model miscalibration and overconfidence can result in serious diagnostic errors, these findings are especially crucial.

The table shows that Model C offers better calibration, increased trustworthiness, and more meaningful confidence estimates in addition to better prediction accuracy. This demonstrates the usefulness of information geometry in a variety of machine learning scenarios and validates its theoretical benefits.

4.5 Ablation Study: Effect of Geometric Terms

To isolate the impact of the geometric components, the study performed an **ablation study** using the MNIST (digits 0–4) and EyePACS datasets. The study varied the training configuration by toggling:

Bayesian posterior (Yes/No)

Natural gradient update (Yes/No)

Fisher-based preconditioning (Yes/No)

Table 3: Ablation Study on Geometric Components (MNIST Subset)

Configuration	Bayesian	Natural Gradient	Fisher Preconditioning	Accuracy (%)	ECE (%)	NLL
Standard SGD (Model A)	✗	✗	✗	92.8	6.2	0.174
Bayesian + Euclidean (Model B)	✓	✗	✗	93.7	3.8	0.140
Bayesian + NG (Model C)	✓	✓	✓	**95.0**	**1.9**	**0.109**
Bayesian + NG, no FIM approx.	✓	✓	✗	94.2	2.7	0.123
Bayesian + FIM only	✓	✗	✓	93.3	3.1	0.132

Using the MNIST (digits 0–4) subset, Table 3 shows an ablation study that separates the effects of geometric components, natural gradient descent and Fisher Information Matrix (FIM) preconditioning within the Bayesian learning framework.

With an expected calibration error (ECE) of 6.2% and a negative log-likelihood (NLL) of 0.174, the baseline model, Standard SGD (Model A), which does not incorporate Bayesian inference or any geometric optimisation, achieves 92.8% accuracy. Accuracy increases to 93.7% when Bayesian inference is added without any geometric components (Bayesian + Euclidean, or Model B), while ECE and NLL drop to 3.8% and 0.140, respectively. This suggests that calibration and predictive fit are improved by posterior estimation alone.

The best results are obtained when Bayesian inference is fully integrated with both natural gradient optimisation and FIM preconditioning (Model C), which reduces ECE and NLL to 1.9% and 0.109, respectively, and pushes accuracy to 95.0%.

Xi'an Shiyou Daxue Xuebao (Ziran Kexue Ban)/
Journal of Xi'an Shiyou University, Natural Sciences Edition
ISSN: 1673-064X
E-Publication: Online Open Access
Vol: 68 Issue 08 | 2025
DOI: 10.5281/zenodo.16925990

This illustrates how probabilistic reasoning and geometric optimisation work in concert to greatly increase model accuracy and confidence.

Two more configurations are taken into consideration in order to further deconstruct these effects. Accuracy is still high at 94.2% when the natural gradient is used without Fisher preconditioning (Bayesian + NG, no FIM approx.), but both ECE (2.7%) and NLL (0.123) are significantly higher than in the fully geometrically-informed model. In contrast, a more modest gain of 93.3%, 3.1% ECE, and 0.132 NLL is obtained when FIM preconditioning is applied without natural gradient updates (Bayesian + FIM only).

According to these results, the incorporation of information geometric components, specifically, the natural gradient in conjunction with the FIM approximation, significantly improves both predictive accuracy and uncertainty reliability, even though Bayesian inference remains a crucial basis for calibrated learning. The theoretical claim that more reliable AI systems result from curvature-aware optimisation in the parameter manifold is strongly supported empirically by the ablation results.

4.6 Extended Ablation Study: Boston Housing Regression Task

The study trained and evaluated five variants of the base model under different combinations of Bayesian inference and geometric awareness. For each, the study report:

Negative Log-Likelihood (NLL): Measures model fit under the predicted distribution.

Root Mean Squared Error (RMSE): Measures point prediction error.

Predictive Standard Deviation ($\hat{\sigma}$): Mean width of model-predicted confidence intervals.

Log Predictive Density (LPD): Higher is better; measures log likelihood of observed outcomes under the predictive distribution.

Table 4: Ablation Results – Boston Housing (Regression)

Configuration	Bayesian	Natural Gradient	Fisher Preconditioning	NLL ↓	RMSE ↓	$\hat{\sigma}$ ↓	LPD ↑
Model A: Standard SGD	✗	✗	✗	2.45	4.82	–	-1.88
Model B: Bayes, Euclidean opt.	✓	✗	✗	2.03	4.21	1.17	-1.24
Model C: Bayes + NG + FIM	✓	✓	✓	**1.88**	**3.96**	**0.97**	**-1.03**
Bayes + NG, no Fisher approx.	✓	✓	✗	1.96	4.12	1.08	-1.13
Bayes + FIM, no NG	✓	✗	✓	1.94	4.07	1.03	-1.11

An ablation study on the Boston Housing regression task is presented in Table 4, which looks at the separate and combined effects of Fisher preconditioning, natural gradient optimisation, and Bayesian inference. Evaluating each component's contribution to uncertainty quality and predictive performance in a continuous-output setting is the aim.

Xi'an Shiyou Daxue Xuebao (Ziran Kexue Ban)/
Journal of Xi'an Shiyou University, Natural Sciences Edition
ISSN: 1673-064X
E-Publication: Online Open Access
Vol: 68 Issue 08 | 2025
DOI: 10.5281/zenodo.16925990

With a negative log-likelihood (NLL) of 2.45 and a root mean squared error (RMSE) of 4.82, the baseline, Model A (Standard SGD), which is devoid of both Bayesian and geometric components, produces the worst results. The predictive standard deviation (σ^σ^) is not reported because the model is a point-estimate and does not provide a way to quantify uncertainty.

Model B, which introduces Bayesian inference without geometric enhancements, enhances predictive reliability and model fit. With a mean predictive standard deviation of 1.17, it permits uncertainty estimation while lowering the NLL to 2.03 and the RMSE to 4.21. Under the uncertainty of the model itself, the corresponding log predictive density (LPD) improves from -1.88 to -1.24, indicating more consistent predictions.

Model C, which combines Bayesian inference with Fisher Information Matrix preconditioning and natural gradient updates, performs the best. This configuration narrows predictive uncertainty ($\hat{\sigma}$ = 0.97 $\hat{\sigma}$ = 0.97), maximises LPD (-1.03), and achieves the lowest NLL of 1.88 and the lowest RMSE of 3.96.

These improvements demonstrate that learning with geometry awareness not only increases accuracy but also sharpens confidence intervals, producing predictions that are more reliable and instructive.

Performance slightly deteriorates when applying natural gradients alone without Fisher preconditioning: NLL rises to 1.96, RMSE falls to 4.12, and $\hat{\sigma}$ expands to 1.08. With an NLL of 1.94 and an RMSE of 4.07, eliminating the natural gradient while keeping Fisher preconditioning likewise produces worse results than Model C.

The findings highlight how information geometry and Bayesian reasoning work in tandem. The combination of curvature-aware preconditioning and natural gradient optimisation through the Fisher metric greatly increases point prediction accuracy and the statistical coherence of uncertainty estimates, whereas Bayesian inference enhances model calibration and predictive variance. This demonstrates how information-geometric learning can be used practically even in low-dimensional, tabular regression problems.

5. IMPLICATIONS FOR AI TRANSPARENCY

Transparency, interpretability, and accountability are more important than ever as AI systems are incorporated more and more into decision-making infrastructures, from financial screening and criminal risk assessment to clinical diagnosis (Rudin, 2019; Lipton, 2018).

This section explains how the suggested framework, Bayesian learning based on information geometry, provides an operationally efficient and statistically sound route to transparent, auditable, and reliable AI systems.

5.1 From Forecasting Capabilities to Probabilistic Responsibility

Predictive accuracy is the primary criterion used to evaluate the majority of cutting-edge deep learning systems. But accuracy isn't enough in high-stakes situations.

Xi'an Shiyou Daxue Xuebao (Ziran Kexue Ban)/
Journal of Xi'an Shiyou University, Natural Sciences Edition
ISSN: 1673-064X
E-Publication: Online Open Access
Vol: 68 Issue 08 | 2025
DOI: 10.5281/zenodo.16925990

Clinicians, regulators, and legislators who make decisions need systems that are:

i. Accurately calibrated: Confidence scores ought to represent actual probabilities.
ii. Uncertainty-aware: Capable of articulating the model's uncertainty.
iii. Interpretable: Able to be comprehended, examined, and questioned.

The findings show that frequentist models (such as SGD-trained networks) frequently exhibit miscalibration, a pathology in which they overestimate their incorrect predictions. A major drawback of black-box models is addressed by the geometrically-aware Bayesian models, which offer both precise predictions and well-calibrated probabilistic confidence (Guo et al., 2017).

5.2 Model Confidence as a Layer of Communication

Models can express epistemic uncertainty, or what the model does not know because of sparse data or contradicting evidence, using Bayesian posterior distributions over parameters. This becomes particularly crucial in:

i. Medical triage: a referral to human specialists may be prompted by a prediction with a high degree of uncertainty.
ii. Autonomous systems: these are those in which human intervention or a fallback behaviour may be triggered by uncertainty.
iii. Fraud detection: this is where cases that are unclear might be marked for human review.

Integrating information geometry improves the stability and effectiveness of the learning dynamics. Confidence intervals are guaranteed to be present and statistically coherent by natural gradient descent, which reflects significant distances in the space of probability distributions (Amari, 1998; Martens, 2020). Because of this, probabilistic interpretability is made possible, and model confidence itself serves as a communication channel between human and machine decision-makers.

5.3 Saliency Alignment and Visual Interpretability

The study found that geometrically-informed Bayesian models improved visual saliency alignment and predictive performance in our medical imaging case study (diabetic retinopathy classification). Grad-CAM visualisations showed that:

i. Model A (frequentist) frequently gave irrelevant regions the wrong attention.
ii. Localised pathology-congruent retinal features in Model C (Bayes + NG) facilitate clinicians' confidence in and validation of model judgements.

In fields where decisions need to be justified and where AI is supposed to support expert judgement rather than replace it, this kind of alignment is essential (Caruana et al., 2015).

Xi'an Shiyou Daxue Xuebao (Ziran Kexue Ban)/
Journal of Xi'an Shiyou University, Natural Sciences Edition
ISSN: 1673-064X
E-Publication: Online Open Access
Vol: 68 Issue 08 | 2025
DOI: 10.5281/zenodo.16925990

5.4 Transparency Beyond Justification: Epistemic Stability and Reproducibility

The findings imply that transparency must also include epistemic robustness, or the system's capacity to generate reliable, consistent results under a variety of training scenarios that are plausible, even though explainability has emerged as a key theme in AI ethics.

An extra layer of inductive bias is introduced by using Fisher-based natural gradients, variational inference, and Bayesian priors, which enhances:

i. Convergence reproducibility across random initialisations.
ii. Stability of parameters, which lowers model behaviour variance.
iii. Smoother predictive landscapes show resilience to adversarial perturbations.

Transparency in this context refers to both interpretability and the consistency of the conclusions the model makes, which increases its viability in practical applications (Doshi-Velez & Kim, 2017).

5.5 Consequences for Regulation and Governance of AI

New AI laws (such as the FDA's guidelines on machine learning in healthcare and the EU's AI Act) increasingly require AI systems to:

i. Document risk and uncertainty.
ii. Encourage decision-making that involves human involvement.
iii. Assure algorithmic responsibility.

One way to meet these needs is through geometrically-aware Bayesian learning, which quantifies uncertainty using posterior variance and predictive entropy.

i. By lowering overfitting, unfair bias brought about by noise or under-represented classes is lessened.
ii. Increasing auditability through testable and validated probabilistic outputs that are clearly defined.

This framework is in line with the larger movement to make algorithmic transparency a governance mechanism as opposed to just a technical one.

5.7 A Philosophical Viewpoint: What Does "Understanding" Mean?

Lastly, the framework raises an important question in scientific AI: Is uncertainty estimation an explanation in and of itself? The study contends that probabilistic transparency, which is based on statistically grounded uncertainty, is a more profound and moral form of explanation, one that recognises the limitations of both data and model than interpretability, which is frequently confused with post hoc visualisation or feature attribution.

Xi'an Shiyou Daxue Xuebao (Ziran Kexue Ban)/
Journal of Xi'an Shiyou University, Natural Sciences Edition
ISSN: 1673-064X
E-Publication: Online Open Access
Vol: 68 Issue 08 | 2025
DOI: 10.5281/zenodo.16925990

6. CONCLUSION

Based on the integration of information geometry and Bayesian inference, this study offered a cohesive, statistically supported framework for transparent and reliable artificial intelligence.

The study showed how geometrically-aware Bayesian learning enhances not only predictive performance but also important aspects of model reliability, interpretability, and epistemic robustness, qualities that are becoming more and more crucial in practical AI applications, through both theoretical explanation and empirical validation.

The study's strategy is based on the understanding that deep learning models, despite their strength, frequently lack mechanisms for representing, calibrating, and explaining uncertainty.

Stochastic gradient descent and other standard optimisation techniques work in a Euclidean parameter space that ignores the model landscape's statistical structure and local curvature.

On the other hand, the framework provides a Riemannian metric to parameter space through the Fisher Information Matrix, allowing for statistically efficient and geometrically coherent updates through natural gradient descent.

At the same time, uncertainty-aware inference using posterior distributions rather than point estimates is made possible by the application of Bayesian learning principles. In addition to improving performance, this dual focus on probability and geometry produces models that can flag epistemic risk, communicate uncertainty, and postpone decision-making in situations that are unclear or involve significant stakes.

Across artificial tasks, tabular regression, image classification, and medical diagnosis, the empirical results consistently favoured the suggested approach (Model C: Bayes + Natural Gradient) over both traditional Bayesian networks (Model B) and standard deep learning (Model A).

In addition to having smoother convergence and better visual interpretability (e.g., saliency map alignment with clinical markers), Model C performed better in terms of accuracy, negative log-likelihood, expected calibration error, and entropy-based uncertainty.

The advantages of our framework went beyond metrics in our case study on diabetic retinopathy. The model served as an explainable and audit-ready assistant to human practitioners in addition to being a predictor by generating posterior-based confidence intervals and clinically-aligned attention heatmaps.

This highlights a larger trend in AI: from systems that aim to surpass humans to those that collaborate, communicate, and provide human-comprehensible justifications for their decisions.

Xi'an Shiyou Daxue Xuebao (Ziran Kexue Ban)/
Journal of Xi'an Shiyou University, Natural Sciences Edition
ISSN: 1673-064X
E-Publication: Online Open Access
Vol: 68 Issue 08 | 2025
DOI: 10.5281/zenodo.16925990

Summary of Contributions

Based on probabilistic modelling and differential geometry, we developed a natural gradient-based Bayesian learning framework.

1. The study combined calibration analysis, entropy estimation, and saliency map interpretability to create an empirical evaluation pipeline that was applied to both classification and regression tasks.
2. The study showed that the Fisher-aware updates enhance both convergence and uncertainty quality through ablation studies that isolate the impact of geometric regularisation.
3. The study connected theoretical accuracy with practical impact by using a case study in medical imaging to demonstrate the framework's practical usefulness.

Future Work

There are still a number of avenues for expansion:

1. Scalability: Scaling to large transformer-based architectures and diffusion models requires effective approximations to the Fisher Information Matrix (e.g., K-FAC, diagonal estimates).
2. Geometry-aware priors: To better align model assumptions with geometric insights, future research could investigate non-Euclidean priors, such as those defined on manifolds or Lie groups.
3. Multi-modal inference: By incorporating this framework into multi-view Bayesian deep learning (such as vision + text + tabular), strong, uncertainty-aware fusion models may be produced.
4. Integration with decision theory: Principled action selection and reinforcement learning may result from integrating geometrically-informed Bayesian learning into a Bayesian decision theory framework.

In the end, this paper makes the case that mathematics is still important for the development of artificial intelligence. We can create intelligent, interpretable, and morally consistent AI systems by expanding upon the fundamental statistical concepts of probability, information, and geometry.

References

1) **Amari, S.** (1998). Natural gradient works efficiently in learning. *Neural Computation*, 10(2), 251–276. https://doi.org/10.1162/089976698300017746

2) **Amari, S., & Nagaoka, H.** (2000). *Methods of Information Geometry*. American Mathematical Society and Oxford University Press.

3) **Blundell, C., Cornebise, J., Kavukcuoglu, K., & Wierstra, D.** (2015). Weight uncertainty in neural networks. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)* (pp. 1613–1622). PMLR.

ISSN: 1673-064X
E-Publication: Online Open Access
Vol: 68 Issue 08 | 2025
DOI: 10.5281/zenodo.16925990

4) **Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N.** (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1721–1730). https://doi.org/10.1145/2783258.2788613

5) **Doshi-Velez, F., & Kim, B.** (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint*, arXiv:1702.08608. https://arxiv.org/abs/1702.08608

6) **Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S.** (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639), 115–118. https://doi.org/10.1038/nature21056

7) **Gal, Y., & Ghahramani, Z.** (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning* (pp. 1050–1059). PMLR.

8) **Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q.** (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning* (pp. 1321–1330). PMLR.

9) **Khan, M. E., Lin, W., Nielsen, D., et al.** (2018). Fast and scalable Bayesian deep learning by weight-perturbation in Adam. In *Bayesian Deep Learning Workshop*, ICML 2018.

10) **LeCun, Y., Bengio, Y., & Hinton, G.** (2015). Deep learning. *Nature*, 521(7553), 436–444. https://doi.org/10.1038/nature14539

11) **Lipton, Z. C.** (2018). The mythos of model interpretability. *Communications of the ACM*, 61(10), 36–43. https://doi.org/10.1145/3233231

12) **MacKay, D. J. C.** (2003). *Information Theory, Inference and Learning Algorithms*. Cambridge University Press.

13) **Martens, J., & Grosse, R.** (2015). Optimizing neural networks with Kronecker-factored approximate curvature. In *Proceedings of the 32nd International Conference on Machine Learning* (pp. 2408–2417). PMLR.

14) **Martens, J.** (2020). New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146), 1–76. http://jmlr.org/papers/v21/19-1016.html

15) **Neal, R. M.** (1995). *Bayesian Learning for Neural Networks* (Doctoral dissertation, University of Toronto). Springer-Verlag.

16) **Pascanu, R., & Bengio, Y.** (2014). Revisiting natural gradient for deep networks. In *International Conference on Learning Representations (ICLR)*.

17) **Rudin, C.** (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x