SEPARATING POSTERIOR GEOMETRY FROM INFERENCE MECHANISMS: A UNIFIED CAYLEY-ORBIT POSTERIOR FOR BAYESIAN NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Bayesian neural networks (BNNs) quantify epistemic uncertainty through posterior distributions over weights. Usually, weight posteriors in BNNs primarily control the scale and statistical dependence of perturbations in an unconstrained weight space, providing a tractable representation of weight uncertainty. However, such posterior parameterizations can suffer from restricted directional exploration and unstable weight structure during sampling. Moreover, posterior construction is often embedded within a particular inference mechanism, making useful posterior geometries difficult to transfer across BNN methods. To address these limitations, we introduce **Unified Cayley-Orbit Posterior (UniCOP)**, a unified posterior constructor separated from the posterior inference mechanism that is used to learn the posterior parameters. Our UniCOP constructs a posterior on a curved orbit by applying a Cayley rotation to a learnable reference weight, preserving the reference weight's structure. This geometry-aware construction better models the epistemic uncertainty through directional exploration while keeping structural properties, supporting predictive stability and robustness. Moreover, our constructor can therefore be employed with both explicit variational and optimizer-induced inference mechanisms. Our theoretical analysis characterizes how the Cayley-orbit construction preserves weight-matrix structure and establishes the geometric validity and versatility of UniCOP. Extensive experiments across multiple BNN methods also demonstrate consistent improvements of UniCOP on uncertainty estimation, robustness to input perturbations, and predictive generalization.

1 INTRODUCTION

Bayesian neural networks (BNNs) quantify epistemic uncertainty through posterior distributions over model weights. Since exact Bayesian inference is generally intractable, variational methods learn tractable approximations to these weight posteriors (Jordan et al., 1999; Blei et al., 2017). Existing posterior families have become increasingly expressive through learned scales, covariance structures, low-rank factors, and normalizing flows (Blundell et al., 2015; Dusenberry et al., 2020). These developments provide flexible ways to represent the magnitude and statistical dependence of weight uncertainty, making variational BNNs applicable at modern neural-network scales.

Despite this progress, existing BNNs have two limitations. The first one is insufficient geometric control over BNN posteriors. Previous approaches often oversimplify posterior distributions to balance computational efficiency and predictive accuracy in high-dimensional parameter spaces. This can restrict directional exploration and hinder reliable epistemic uncertainty estimation. More expressive posterior structures allow more flexible exploration, but fail to preserve geometric properties of weights, such as norms and singular values (Louizos & Welling, 2016), potentially compromising stability during posterior learning. Specifically in a rotation orbit, a narrow distribution covers only a small portion of the orbit, whereas a broad distribution places probability mass outside it (Rezende & Mohamed, 2015). This motivates a posterior constructor that enables directional exploration along curved rotation orbits while preserving weight structure.

The second limitation concerns reuse across posterior inference mechanisms. Explicit variational mechanisms like Bayes by Backprop (BBB) (Blundell et al., 2015) directly learn the parameters of

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

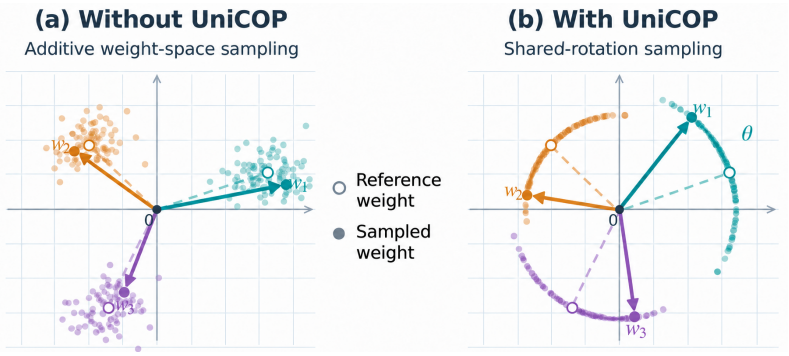


Figure 1: **Comparison of weight samples with and without UniCOP constructor.** Colors identify three rows of a weight matrix. Hollow circles mark the reference row endpoints, filled circles highlight one joint weight sample, and faint dots show endpoints across repeated samples. Dashed and solid arrows connect the origin to the reference and sampled endpoints, respectively. (a) Without UniCOP, conventional weight-space sampling directly perturbs the rows, allowing their lengths and pairwise angles to change. (b) With UniCOP, each draw applies a shared rotation to all reference rows, varying their orientations while preserving their lengths and pairwise angles.

a weight posterior, whereas optimizer-induced inference such as VOGN (Osawa et al., 2019) and IVON (Shen et al., 2024) derive posterior scales from curvature or precision statistics maintained during training (Khan et al., 2018). In both cases, posterior design is embedded within the corresponding stochastic layer or inference procedure, making it difficult to apply the same posterior geometric design consistently across BNN inference mechanisms.

To address both limitations, we propose **Unified Cayley-Orbit Posterior (UniCOP)**, a unified posterior constructor that is separated from the posterior inference mechanism. The constructor defines how weights are sampled from the approximate posterior, while the inference mechanism specifies how the posterior-defining parameters are learned from data. This separation allows the same constructor to be employed with both explicit variational and optimizer-induced inference mechanisms. As for posterior construction, we design the UniCOP constructor from a geometric perspective, to better model epistemic uncertainty through weights' directional exploration with structural preservation. Specifically, it transforms an auxiliary Gaussian matrix into a rotation matrix through skew-symmetric extraction and the Cayley transform (Helfrich et al., 2018). Applying this rotation to a learnable reference weight generates samples on its rotation orbit, varying weight orientation while preserving the reference weight's row Gram matrix, Frobenius norm, and singular values.

Our theory establishes the validity of latent-space variational learning and characterizes how UniCOP represents uncertainty in weight orientation while preserving matrix structure. It shows that UniCOP represents orbit-aligned uncertainty without placing probability mass outside the rotation orbit, and that the constructor's geometry can change predictive uncertainty under the same latent KL. Mechanistic studies with BBB and IVON isolate these properties under explicit and optimizer-induced inference mechanisms. The main experiments further demonstrate improvements in corruption robustness, epistemic out-of-distribution ranking, and class-incremental transfer.

To sum up, our contributions are as follows: a) We are the first to propose a unified framework for posterior construction, enabling the same posterior geometry to be employed with different posterior inference mechanisms. b) We introduce UniCOP, which maps a Gaussian latent matrix through skew-symmetric extraction and a Cayley transform to produce curved weight posteriors with exact structural invariants. c) Our theoretical analysis shows how the Cayley-orbit construction preserves weight-matrix structure and establishes the geometric validity and versatility of UniCOP. d) Extensive experiments across posterior inference mechanisms show consistent improvements in-domain performance, out-of-distribution generalization, and downstream task.

2 Related Work

Explicit Variational Posteriors Explicit variational methods represent posterior parameters as trainable model variables and optimize them through a variational objective. Bayes by Backprop

introduces reparameterized Gaussian weight posteriors with learnable means and scales (Blundell et al., 2015). Local reparameterization moves sampling from weights to activations to reduce gradient variance (Kingma et al., 2015), while Flipout decorrelates perturbations across examples within a mini-batch (Wen et al., 2018). Subsequent work improves posterior expressiveness through matrix-variate Gaussians that capture layer-wise correlations (Louizos & Welling, 2016), multiplicative normalizing flows that introduce flexible dependencies (Louizos & Welling, 2017), radial posteriors that avoid undesirable high-dimensional Gaussian behavior (Farquhar et al., 2020), and rank-one factors that provide parameter-efficient structured uncertainty (Dusenberry et al., 2020).

Optimizer-Induced Approximate Posteriors Optimizer-induced methods construct approximate posteriors from statistics accumulated during learning. Natural-gradient variational inference and the Bayesian learning rule interpret gradient and curvature updates as inference in distribution space (Amari, 1998; Khan & Rue, 2023). VOGN uses online curvature estimates to update a local Gaussian approximation (Khan et al., 2018; Osawa et al., 2019), while IVON provides an Adam-like implementation whose precision statistics determine both optimization steps and posterior perturbations (Shen et al., 2024). Related approaches obtain uncertainty from noisy trajectories or local solution geometry: SGLD injects calibrated update noise to approximate posterior sampling (Welling & Teh, 2011), stationary-distribution analyses connect stochastic optimization to approximate Bayesian inference (Mandt et al., 2017), SWAG fits a low-rank-plus-diagonal Gaussian to SGD iterates (Maddox et al., 2019), and Laplace methods use Hessian structure around a mode (Daxberger et al., 2021).

Structured Posteriors and Geometric Parameterizations Structured variational posteriors use matrix covariances, low-rank factors, auxiliary variables, or flows to capture dependencies omitted by mean-field approximations (Louizos & Welling, 2016; 2017; Dusenberry et al., 2020). These approaches enrich statistical dependence or change the inference space, but do not explicitly determine the structure-preserving directions along which weight uncertainty should propagate. Information geometry and Riemannian optimization address non-Euclidean parameter spaces (Amari, 1998; Absil et al., 2008), while Stiefel and orthogonal optimization preserve matrix structure in deterministic networks (Huang et al., 2018). Manifold trivializations enable Euclidean optimization under such constraints (Lezcano-Casado, 2019), and spectral normalization controls operator norms (Miyato et al., 2018). Probabilistic models have also employed rotation-group distributions and Cayley parameterizations to encode geometric structure (Gillespie et al., 2024).

3 METHODOLOGY

3.1 POSTERIOR DECOMPOSITION FOR BNNS

In BNNs, each randomized layer requires both a posterior over its weights and an inference mechanism for learning that posterior from data. Existing methods often implement these two roles together. We instead separate them through a **posterior constructor**, which specifies how weight samples are constructed, and a **posterior inference mechanism**, which specifies how the posterior parameters are learned.

Given posterior parameters ϕ and base noise ϵ, the posterior constructor produces a weight sample $W = G(\phi, \epsilon)$ (Krueger et al., 2017). For UniCOP, $\phi = (W^\circ, M, S)$ consists of a reference weight matrix and the parameters of a latent Gaussian distribution. Different BNN inference mechanisms can learn these parameters in different ways while sharing the same constructor. Explicit variational methods learn distribution parameters directly, whereas optimizer-induced methods derive sampling scales from precision statistics accumulated during training.

This separation allows UniCOP to modify posterior construction without replacing the posterior inference mechanism. Specific objectives, gradient updates, and auxiliary statistics remain part of the corresponding inference procedure. We refer to this as the **unified posterior construction**.

3.2 UNICOP POSTERIOR CONSTRUCTOR

The preceding separation exposes the posterior constructor as the component through which posterior construction can be designed independently of the inference mechanism. We instantiate this

interface with the Unified Cayley-Orbit Posterior (UniCOP). For each randomized BNN weight matrix, the posterior constructor first samples a full latent matrix,

$$R = M + \exp(S) \odot \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I), \tag{1}$$

where M and S are the elementwise latent mean and log standard deviation respectively. It then extracts the skew-symmetric component $A(R) = \frac{1}{2}(R - R^\top)$ and maps it to an orthogonal matrix through the Cayley transform $Q(R) = (I - A(R))(I + A(R))^{-1}$ (Helfrich et al., 2018; Lezcano-Casado, 2019). Then the weight is sampled from

$$W = f_{W^\circ}(R) = W^\circ Q(R)^\top, \tag{2}$$

where W° is a learnable but non-sampled reference weight. The set $\mathcal{O}(W^\circ) = \{W^\circ Q^\top : Q \in \mathrm{SO}(d)\}$ is the rotation orbit of W°. Because $A(R)$ is skew-symmetric, $Q(R)$ is orthogonal. Consequently, every sampled weight satisfies $WW^\top = W^\circ W^{\circ\top}$, $|W|_F = |W^\circ|_F$, $\sigma_i(W) = \sigma_i(W^\circ)$. The constructor therefore varies the orientation of W° while preserving its row Gram matrix, Frobenius norm and singular values. We refer to this property as **structure-preserving control** with the detailed explanations in section 4.2.

The corresponding classical Gaussian posterior samples weight generally as

$$W_0 = \mu + \exp(S) \odot \epsilon. \tag{3}$$

In comparison, the classical diagonal Gaussian represents uncertainty through a linear cloud in weight space, whereas the UniCOP constructor maps a diagonal latent Gaussian to a curved and correlated distribution on the rotation orbit $\mathcal{O}(W^\circ)$. UniCOP covers the relevant curved variation without allocating probability mass to off-orbit directions. We call this property **geometric representation efficiency** and the details are in section 4.3.

Since UniCOP constructor is fully differentiable, its posterior parameters and reference weight can be learned through standard backpropagation using the original inference mechanism with its original update rules and method-specific statistics. UniCOP thus changes how BNN weight posteriors are constructed without modifying how they are learned.

3.3 Latent-Space Variational Inference

Having defined the UniCOP constructor, we instantiate an inference mechanism to learn its posterior parameters. We place a prior and an approximate posterior over the auxiliary latent matrix R:

$$p_R = \mathcal{N}(0, \sigma_0^2 I_{n_R}), \qquad q_R = \mathcal{N}(M, \mathrm{Diag}(\exp(2S))), \tag{4}$$

where n_R denotes the dimension of R. Both distributions are defined over the full latent matrix before UniCOP applies skew-symmetric extraction and the Cayley transform. A sample from the latent posterior can equivalently be written as

$$R = M + \exp(S) \odot \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I). \tag{5}$$

The reference weight W° is optimized through the expected data term, while the posterior constructor f_{W° maps the latent distributions to weight-space distributions. Thus, $q_W = (f_{W^\circ})_\# q_R$ is the approximate posterior over generated weights obtained by sampling $R \sim q_R$ and setting $W = f_{W^\circ}(R)$, while $p_W = (f_{W^\circ})_\# p_R$ is the corresponding weight prior obtained from $R \sim p_R$. The notation $(f_{W^\circ})_\#$ denotes the pushforward through f_{W°.

Then the posterior-defining parameters W°, M and S are learned by minimizing the negative latent-space evidence lower bound:

$$\mathcal{J}_{\mathrm{G}}(W^\circ, M, S) = \mathbb{E}_{R \sim q_R}\left[-\log p(\mathcal{D} \mid f_{W^\circ}(R))\right] + \mathrm{KL}(q_R \| p_R). \tag{6}$$

For the diagonal Gaussian q_R and isotropic Gaussian p_R in equation 4, the KL divergence is

$$\mathrm{KL}(q_R \| p_R) = \frac{1}{2} \sum_{i=1}^{n_R} \left[\frac{m_i^2 + \exp(2s_i)}{\sigma_0^2} - 1 + 2\log \sigma_0 - 2s_i \right], \tag{7}$$

where m_i and $\exp(s_i)$ are the mean and standard deviation of the i-th vectorized entry of R. The variational objective therefore remains tractable in the Gaussian latent space even though the UniCOP constructor induces a curved and potentially singular posterior in weight space.

Under the isotropic latent prior, different allocations of the same covariance spectrum across latent directions have the same KL value, whereas the UniCOP constructor and the model Jacobian determine how those directions affect function-space predictions. We refer to this property as **fixed-KL function-space allocation** with its detailed explanations in 4.4.

4 THEORY

4.1 VALIDITY OF THE LATENT OBJECTIVE

The many-to-one constructor can induce a curved or singular weight posterior, for which a weight-space density need not exist. We therefore establish the variational objective directly in latent space.

Theorem 1 (**Validity of latent-space variational learning**). *Fix W° and a measurable constructor f_{W°. Let p_R and q_R be latent distributions satisfying $q_R \ll p_R$ and $\mathrm{KL}(q_R \| p_R) < \infty$. Define $p_W = (f_{W^\circ})_{\#} p_R$, $q_W = (f_{W^\circ})_{\#} q_R$, and $p(\mathcal{D} \mid W^\circ) = \int p(\mathcal{D} \mid f_{W^\circ}(R)) p_R(\mathrm{d}R)$. Assume that this evidence is finite and positive and that the expected log-likelihood exists.*

The latent objective is a valid lower bound, and its KL term upper-bounds the weight-space KL:

$$\log p(\mathcal{D} \mid W^\circ) \geq \mathbb{E}_{R \sim q_R}[\log p(\mathcal{D} \mid f_{W^\circ}(R))] - \mathrm{KL}(q_R \| p_R), \tag{8}$$

$$\mathrm{KL}(q_W \| p_W) \leq \mathrm{KL}(q_R \| p_R). \tag{9}$$

The first inequality certifies the optimized latent objective as a valid ELBO. The second shows that its tractable latent KL upper-bounds the divergence between the induced weight distributions. UniCOP therefore supports valid variational learning without requiring a weight-space density. The proof is given in Appendix A.1.1.

4.2 STRUCTURE-PRESERVING CONTROL

Variational validity alone does not identify which weight directions carry uncertainty. This result characterizes the exact support and local directions produced by UniCOP.

Theorem 2 (**Exact structure-preserving support**). *Let $W^\circ \in \mathbb{R}^{m \times d}$, let $\mathrm{SO}(d)$ be the group of $d \times d$ rotations, and define $\mathcal{O}(W^\circ) = \{W^\circ Q^\top : Q \in \mathrm{SO}(d)\}$. For $R \in \mathbb{R}^{d \times d}$, define $A(R) = \frac{1}{2}(R - R^\top)$, $Q(R) = (I - A(R))(I + A(R))^{-1}$, and $f_{W^\circ}(R) = W^\circ Q(R)^\top$. For any $M \in \mathbb{R}^{d \times d}$, let $W_M = f_{W^\circ}(M)$, let $\mathrm{D}f_{W^\circ}(M)$ be the constructor differential, and let $T_{W_M}\mathcal{O}(W^\circ)$ be the orbit tangent space at W_M. Let $\sigma_i(U)$ denote the i-th singular value of U.*

For every R and every singular-value index i, the posterior remains on the orbit, preserves the reference structure, and spans exactly the orbit's local directions:

$$\begin{aligned} W = f_{W^\circ}(R) \in \mathcal{O}(W^\circ), \qquad & WW^\top = W^\circ W^{\circ\top}, \qquad \|W\|_F = \|W^\circ\|_F, \\ \sigma_i(W) = \sigma_i(W^\circ), \qquad & \mathrm{Im}\big(\mathrm{D}f_{W^\circ}(M)\big) = T_{W_M}\mathcal{O}(W^\circ). \end{aligned} \tag{10}$$

Every posterior sample remains on the same rotation orbit and exactly preserves the row Gram matrix, Frobenius norm, and singular values. The tangent-space equality further shows that the constructor covers all locally available orbit directions and no off-orbit direction. This is the precise meaning of **structure-preserving control**. The proof is given in Appendix A.1.2.

4.3 GEOMETRIC REPRESENTATION EFFICIENCY

Structure preservation becomes a representation advantage when the relevant uncertainty follows the orbit. This result compares UniCOP with a classical diagonal Gaussian after matching their useful first-order variation along the orbit.

Theorem 3 (**Zero off-orbit leakage at matched tangent variation**). *Let $W_M = f_{W^\circ}(M)$ and assume that $\mathcal{O}(W^\circ)$ is a proper lower-dimensional orbit. Let $P_T, P_N \in \mathbb{R}^{md \times md}$ be the orthogonal projections onto its vectorized tangent and normal spaces at W_M, and define $d_{\mathcal{O}}(U) = \inf_{V \in \mathcal{O}(W^\circ)} \|U - V\|_F$. Write $w_M = \mathrm{vec}(W_M)$ and let $\mathrm{mat}_{m \times d}$ be the inverse reshape of* vec. *For diagonal $D \in \mathbb{R}^{md \times md}$ with $D \succ 0$ and $\xi \sim \mathcal{N}(0, D)$, define $W_{\mathrm{E}}(t) = \mathrm{mat}_{m \times d}(w_M + t\xi)$. For a random $Z \in \mathbb{R}^{d \times d}$ satisfying $\mathbb{E}\|Z\|_F^4 < \infty$, define $W_{\mathrm{G}}(s) = f_{W^\circ}(M + sZ)$. Assume $c_{\mathrm{E}} = \mathrm{tr}(P_T D) > 0$ and $c_{\mathrm{G}} = \mathbb{E}\|\mathrm{D}f_{W^\circ}(M)[Z]\|_F^2 > 0$, and set $s(t) = t\sqrt{c_{\mathrm{E}}/c_{\mathrm{G}}}$.*

As $t \to 0$, the two constructions have matched leading tangent variation, but only the classical Gaussian places mass away from the orbit:

$$\mathbb{E}\, d_{\mathcal{O}}(W_{\mathrm{E}}(t))^2 = t^2 \operatorname{tr}(P_N D) + O(t^3), \tag{11}$$

$$d_{\mathcal{O}}(W_{\mathrm{G}}(s(t))) = 0 \quad \textit{a.s.}, \tag{12}$$

$$\mathbb{E}\|P_T \operatorname{vec}(W_{\mathrm{G}}(s(t)) - W_M)\|_2^2 = \mathbb{E}\|P_T \operatorname{vec}(W_{\mathrm{E}}(t) - W_M)\|_2^2 + O(t^3). \tag{13}$$

Because $D \succ 0$ and the proper orbit has a nontrivial normal space, the first quantity is strictly positive for all sufficiently small $t > 0$.

At the same leading variation along the orbit, the classical Gaussian allocates nonzero variation to normal directions, whereas UniCOP has exactly zero off-orbit leakage. UniCOP therefore concentrates its posterior capacity on the structure-preserving directions relevant to an orbit-aligned target. This is **geometric representation efficiency**. The proof is given in Appendix A.1.3.

4.4 Fixed-KL Function-Space Allocation

Equal latent KL does not imply equal predictive uncertainty because the constructor maps latent directions to function space differently. This result isolates that allocation effect while holding the KL budget fixed.

Theorem 4 (Fixed-KL function-space allocation). *Let $g_X(R)$ be the vector of model outputs on inputs X when the weights are $f_{W^\circ}(R)$. Let $m = \operatorname{vec}(M) \in \mathbb{R}^{n_R}$ and $J_X = \nabla_{\operatorname{vec}(R)} g_X(M)$. For diagonal $D \succ 0$ and a permutation matrix Π, define $\widetilde{D} = \Pi D \Pi^\top$, $q_t = \mathcal{N}(m, t^2 D)$, and $\widetilde{q}_t = \mathcal{N}(m, t^2 \widetilde{D})$. Draw $r_t \sim q_t$ and $\widetilde{r}_t \sim \widetilde{q}_t$, and define $R_t = \operatorname{mat}_{d\times d}(r_t)$ and $\widetilde{R}_t = \operatorname{mat}_{d\times d}(\widetilde{r}_t)$. Let the prior be $p_R = \mathcal{N}(0, \sigma_0^2 I_{n_R})$, and assume that g_X is twice continuously differentiable with bounded second derivative.*

As $t \to 0$, permuting the latent variances preserves the KL exactly but changes the leading function-space covariance according to the output Jacobian:

$$\mathrm{KL}(q_t \| p_R) = \mathrm{KL}(\widetilde{q}_t \| p_R), \tag{14}$$

$$\operatorname{Cov}[g_X(R_t)] - \operatorname{Cov}[g_X(\widetilde{R}_t)] = t^2 J_X (D - \widetilde{D}) J_X^\top + O(t^3).$$

Permuting the same latent variances preserves the KL exactly, yet changes the leading output covariance whenever $J_X(D - \widetilde{D})J_X^\top \neq 0$. The KL controls the total latent budget, while the constructor and model Jacobians determine where that budget appears in function space. The proof is given in Appendix A.1.4.

5 Instantiation and Mechanistic Evidence

5.1 Explicit Variational Inference: BBB+UniCOP

BBB directly learns a mean-field Gaussian posterior in weight space. BBB+UniCOP retains the same explicit variational inference mechanism but replaces this posterior construction with the UniCOP constructor, which maps a latent Gaussian onto the rotation orbit $\mathcal{O}(W^\circ)$:

$$W_{\mathrm{BBB}} = \mu + \sigma \odot \epsilon, \qquad W_{\mathrm{BBB+UniCOP}} = f_{W^\circ}(M + \exp(S) \odot \epsilon)\,. \tag{15}$$

We construct 20 paired, bias-free binary-classification problems using a fixed two-dimensional reference weight and 28 input directions. The detailed task construction, evaluation metrics, and quantitative results are provided in Appendix A.3.1.

Table 1 further shows that UniCOP reduces predictive KL by $69\times$, angle-W_2 by $4.5\times$, and weight energy distance by $2.3\times$. These results verify the **geometric representation efficiency** predicted by Theorem 3.

Table 1: **Orbit-aligned BBB mechanism.** UniCOP reduces predictive KL by $69\times$, angle-W_2 by $4.5\times$, and weight energy distance by $2.3\times$, indicating closer agreement with the reference posterior or predictive.

Metric	Default	**UniCOP**
Predictive KL ↓	0.015066	0.000218
Angle-W_2 ↓	0.230928	0.051800
Weight energy distance ↓	1.131163	0.488771
Expected dense NLL ↓	0.536515	0.521667

Table 2: **IVON posterior-construction control at matched latent KL.** Both constructions share $q_{t,\alpha}$ at each model–budget point. Entries are absolute means over five models and seven budgets; arrows indicate the preferred direction.

Metric	Euclidean	**UniCOP**
Row-Gram drift ↓	1.26883	1.58×10^{-7}
MI-AUROC ↑	0.89584	0.90061
MI-AURC ↓	0.01967	0.01898
JSD-AUROC ↑	0.89917	0.90127

5.2 Optimizer-Induced Inference: IVON+UniCOP

IVON determines posterior scales from optimizer-maintained precision statistics. IVON+UniCOP retains this inference mechanism and constructs weight samples from the resulting latent Gaussian:

$$S_t = -\frac{1}{2}\log[N_{\text{eff}}(\pi_t + \lambda_A)], \qquad z_t = M_t + \exp(S_t)\odot\epsilon, \qquad W_t = W_t^{\circ}Q(R_t)^{\top}, \tag{16}$$

where M_t is the latent mean, π_t is the maintained precision, N_{eff} is the effective sample size, $\lambda_A = (N_{\text{eff}}\sigma_0^2)^{-1}$ is the prior contribution, and R_t is the reshaped z_t. Only latent parameters are sampled; the reference weight and other deterministic parameters remain point estimates.

To isolate posterior construction, we compare UniCOP to a default Euclidean control with the same center prediction and matched first-order perturbation energy. We evaluate five independently trained IVON models on CIFAR-10 at seven latent KL budgets by rescaling the learnable posterior:

$$q_{t,\alpha}(z) = \mathcal{N}\big(M_t, \alpha^2\,\mathrm{Diag}(\exp(2S_t))\big), \qquad \kappa_{t,\alpha} = \mathrm{KL}(q_{t,\alpha}\|p_z)/d_z, \tag{17}$$

where $\alpha > 0$ scales the learnable standard deviations, p_z is the latent prior over the vectorized variable z, and d_z is its dimension. At each model-budget point, both constructions share this latent posterior, noise draws, reference weight, and KL budget. We assess structure preservation through row-Gram drift and rank the shared center predictor's errors using mutual information (MI) (Houlsby et al., 2011) and pairwise Jensen-Shannon divergence (JSD) (Lin, 1991).

Table 2 shows that UniCOP reduces row-Gram drift by approximately seven orders of magnitude while improving error ranking and lowering MI-AURC. These results demonstrate **structure-preserving posterior directions** and support **fixed-KL function-space allocation**. Control definitions, evaluation details, and full results appear in Appendix A.3.2.

6 Experiments

6.1 In-Domain Prediction

We employ UniCOP with BBB, IVON, Rank-1 BNN (Dusenberry et al., 2020), and VOGN on CIFAR-10 and CIFAR-100 to test the in-domain performance across posterior inference mechanisms. Other baseline models include Deterministic DNN (He et al., 2016), BatchEnsemble (Wen et al., 2020), MC Dropout (Gal & Ghahramani, 2016), SWAG (Maddox et al., 2019), Variational Dropout (VD) (Kingma et al., 2015) and Radial BNN (Farquhar et al., 2020). Table 3 shows that IVON+UniCOP raises accuracy from 91.74% to 92.87% on CIFAR-10 and from 64.31% to 64.67% on CIFAR-100, achieving the SOTA. These results establish that UniCOP can be integrated with distinct posterior inference mechanisms while maintaining competitive clean-data prediction.

6.2 Discrete OOD Detection

We train models on CIFAR-10 and evaluate them on SVHN (Netzer et al., 2011) and Flowers102 (Nilsback & Zisserman, 2008) using the maximum posterior-predictive class probability as the confidence score (Hendrycks & Gimpel, 2017). As shown in Table 4, BBB+UniCOP raises AUROC from 81.35% to 82.76% on SVHN, while IVON+UniCOP lowers FPR95 from 77.07% to 73.96%. All pairings improve AUROC, FPR95 and detection error on both OOD datasets, indicating that UniCOP improves OOD separation under different posterior inference mechanisms.

Table 3: **In-domain prediction** on CIFAR-10 and CIFAR-100 with BNNs on ResNet-20 backbone.

Dataset	Method	Accuracy(%) ↑	NLL ↓	ECE(%) ↓	Brier(%) ↓
CIFAR-10	Deterministic	92.13 ± 0.15	0.299 ± 0.0088	3.28 ± 0.20	14.00 ± 0.18
	BatchEnsemble	89.66 ± 0.14	0.470 ± 0.0036	5.85 ± 0.44	20.07 ± 0.64
	MC-dropout	89.27 ± 0.20	0.300 ± 0.0028	2.75 ± 0.16	14.38 ± 0.17
	SWAG	91.82 ± 0.24	0.296 ± 0.0090	2.70 ± 0.23	12.91 ± 0.34
	VD	90.53 ± 0.23	0.389 ± 0.0127	5.63 ± 0.25	15.13 ± 0.37
	Radial BNN	89.70 ± 0.21	0.404 ± 0.107	4.98 ± 0.39	13.97 ± 0.23
	BBB	88.22 ± 0.45	0.467 ± 0.0209	5.91 ± 0.38	19.39 ± 0.72
	BBB+**UniCOP**	88.90 ± 0.21	0.436 ± 0.0214	5.73 ± 0.34	18.90 ± 0.49
	Rank-1 BNN	90.04 ± 0.18	0.302 ± 0.0044	3.59 ± 0.35	12.82 ± 0.16
	Rank-1 BNN+**UniCOP**	91.56 ± 0.22	0.295 ± 0.0072	3.09 ± 0.48	14.44 ± 0.31
	VOGN	90.98 ± 0.36	0.325 ± 0.0139	4.10 ± 0.31	14.65 ± 0.55
	VOGN+**UniCOP**	91.30 ± 0.40	0.319 ± 0.0099	3.96 ± 0.32	14.53 ± 0.49
	IVON	91.74 ± 0.28	0.297 ± 0.0141	4.36 ± 0.29	12.80 ± 0.50
	IVON+**UniCOP**	92.87 ± 0.19	0.288 ± 0.0055	4.10 ± 0.15	12.64 ± 0.14
CIFAR-100	Deterministic	63.95 ± 0.32	1.294 ± 0.0147	5.14 ± 0.15	49.03 ± 0.50
	SWAG	64.22 ± 0.54	1.259 ± 0.0207	6.01 ± 0.88	48.78 ± 0.54
	BBB	61.26 ± 1.98	1.565 ± 0.0767	9.03 ± 0.40	62.04 ± 1.20
	BBB+**UniCOP**	62.15 ± 0.98	1.525 ± 0.0623	15.99 ± 0.74	60.77 ± 1.08
	Rank-1 BNN	62.56 ± 0.64	1.478 ± 0.0240	12.03 ± 0.53	53.75 ± 0.60
	Rank-1 BNN+**UniCOP**	62.50 ± 0.23	1.597 ± 0.0065	11.77 ± 0.51	57.22 ± 0.04
	VOGN	61.69 ± 0.53	1.555 ± 0.0242	9.81 ± 0.45	54.63 ± 0.79
	VOGN+**UniCOP**	61.97 ± 0.71	1.551 ± 0.0265	9.77 ± 0.65	54.28 ± 0.69
	IVON	64.31 ± 0.42	1.364 ± 0.0152	10.29 ± 0.45	48.97 ± 0.38
	IVON+**UniCOP**	64.67 ± 0.40	1.351 ± 0.0236	10.21 ± 0.49	48.46 ± 0.56

Table 4: **Discrete OOD detection from CIFAR-10.** Values are mean ± sample SD over three matched seeds. FPR95 is measured at the first ROC operating point that attains at least 95% in-domain TPR.

OOD dataset	Method	AUROC(%) ↑	AUPR-out(%) ↑	FPR95(%) ↓	Detection error(%) ↓
Flowers102	BBB	82.61 ± 0.87	69.69 ± 1.65	75.95 ± 1.79	23.97 ± 0.49
	BBB+**UniCOP**	83.50 ± 0.89	70.68 ± 1.76	75.62 ± 1.96	22.90 ± 0.73
	IVON	87.24 ± 0.58	77.24 ± 0.60	64.91 ± 0.47	19.79 ± 0.57
	IVON+**UniCOP**	87.99 ± 1.01	78.56 ± 1.53	63.32 ± 3.07	18.96 ± 0.94
SVHN	BBB	81.35 ± 0.94	88.35 ± 0.66	82.95 ± 1.56	24.16 ± 0.76
	BBB+**UniCOP**	82.76 ± 0.45	89.44 ± 0.43	81.33 ± 1.27	22.83 ± 0.63
	IVON	86.06 ± 0.38	91.21 ± 0.27	77.07 ± 1.00	19.19 ± 0.36
	IVON+**UniCOP**	87.28 ± 0.61	91.96 ± 0.28	73.96 ± 1.39	18.03 ± 0.78

6.3 Continuous Distribution Shift

We evaluate BBB and IVON on CIFAR-10-Corrupted and CIFAR-100-Corrupted (Hendrycks & Dietterich, 2019) across several corruption severities. Figure 2 shows higher accuracy and lower NLL with UniCOP as corruption severity increases from severity 0 to 5. Models with UniCOP also achieve lower E-AURC and selective risk, showing that posterior-predictive entropy more effectively identifies unreliable predictions. The gains in both prediction and selective risk indicate that UniCOP remains useful under continuous distribution shift.

6.4 Downstream Transfer to CLAP4CLIP

We further investigate whether UniCOP can transfer beyond BNN weights to feature-space posteriors in downstream tasks. We integrate UniCOP into CLAP4CLIP (Jha et al., 2024) and evaluate it on CIFAR-100 over multiple incremental sessions and different sequences in which classes are introduced. At each session, learned classes are in-domain and future classes are OOD. CLAP4CLIP constructs a mean-field Gaussian posterior over CLIP text features. We replace this posterior with UniCOP while retaining the backbone, replay strategy, training schedule, and evaluation protocol.

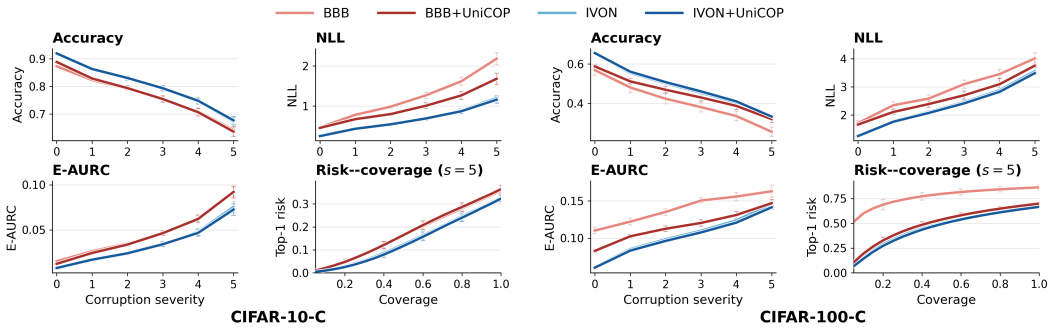


Figure 2: **Continuous corruption robustness on CIFAR-10-C and CIFAR-100-C.** Accuracy, NLL and E-AURC are computed after pooling all 15 corruptions at each severity; the final panel shows entropy-based selective risk at severity 5. Lines and capped error bars are mean ± sample SD over three seeds; lower is better except for accuracy.

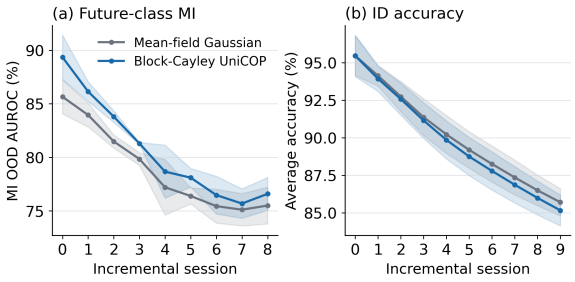


Figure 3: **CLAP4CLIP transfer.** Lines show mean ± population standard deviation over three held-out class orders. (a) With the MI score frozen before evaluation, Block-Cayley UniCOP improves future-class OOD ranking throughout the incremental sequence. (b) Both maintain closely matched in-domain accuracy.

Metric	Mean-field Gaussian	**UniCOP**
ID, forgetting, and calibration (session 9)		
Acc. avg ↑	85.70 ± 0.91	85.16 ± 1.03
Acc. last ↑	78.55 ± 0.12	78.20 ± 0.21
BWT ↑	-10.65 ± 0.14	-10.70 ± 0.06
ECE ↓	21.21 ± 0.16	21.56 ± 0.19
Future-class OOD: frozen MI score (sessions 0–8)		
AUROC ↑	78.95 ± 0.97	80.68 ± 0.97
AUPR ↑	73.47 ± 0.68	79.75 ± 1.01
FPR95 ↓	73.71 ± 1.42	71.35 ± 1.08

Table 5: **Final CLAP4CLIP comparison.** Results are reported as mean ± population standard deviation over three held-out class orders, in percent. UniCOP improves future-class OOD detection while maintaining comparable in-domain classification performance on learned classes.

We evaluate classification and forgetting on learned classes and use posterior mutual information to detect future classes.

Figure 3 and Table 5 show that UniCOP consistently improves future-class OOD ranking across all three orders while maintaining comparable in-domain performance. UniCOP increases AUROC from 78.95% to 80.68% and AUPR from 73.47% to 79.75%, while reducing FPR95 from 73.71% to 71.35%. Its final average accuracy is 85.16% versus 85.70% for the Gaussian baseline, with similar forgetting. Together with the block-energy results in Appendix A.4, these gains demonstrate that UniCOP preserves feature geometry and improves epistemic detection of future classes.

7 Conclusion

We introduced **Unified Cayley-Orbit Posterior (UniCOP)**, a unified posterior constructor separated from the posterior inference mechanism which is used to learn the posterior parameters. UniCOP applies Cayley rotations to a learned reference weight, constructing a posterior on a curved orbit while preserving its row Gram matrix, Frobenius norm, and singular values. This constructor can be employed with both explicit variational and optimizer-induced inference mechanisms. Our theoretical analysis establishes the validity of the latent-space objective and explains how Cayley-orbit geometry supports directional exploration while preserving weight-matrix structure. Mechanistic studies verify these properties, and experiments across multiple BNN methods demonstrate consistent improvements in uncertainty estimation, robustness to input perturbations, and predictive generalization. These results establish posterior geometry as a unified design of posterior construction for improving uncertainty estimation and predictive generalization in Bayesian neural networks.

AI USE STATEMENT

In this work, we used generative AI tools to assist in writing proofs and translating research text. We did not use these tools for other tasks requiring disclosure, or those tasks were not applicable to this work. Additionally, we used generative AI tools to create or revise scientific figures, summarize and analyze existing literature, search for information, improve the readability of the manuscript, format references, and draft parts of the paper. The authors reviewed the AI-assisted material, checked the proofs and translations, and verified literature claims and references against the original sources. We take full responsibility for the final content of this work, including text, claims, and figures produced with the aid of generative AI.

REFERENCES

P.-A. Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, NJ, 2008. doi: 10.1515/9781400830244.

Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998. doi: 10.1162/089976698300017746.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017. doi: 10.1080/01621459.2017.1285773.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1613–1622. PMLR, 2015.

Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace redux—effortless Bayesian deep learning. In *Advances in Neural Information Processing Systems*, volume 34. Curran Associates, Inc., 2021.

Michael Dusenberry, Ghassen Jerfel, Yeming Wen, Yian Ma, Jasper Snoek, Katherine Heller, Balaji Lakshminarayanan, and Dustin Tran. Efficient and scalable Bayesian neural nets with rank-1 factors. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2782–2792. PMLR, 2020.

Sebastian Farquhar, Michael A. Osborne, and Yarin Gal. Radial Bayesian neural networks: Beyond discrete support in large-scale Bayesian deep learning. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1352–1362. PMLR, 2020.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1050–1059, 2016.

Patrick Gillespie, Layal Bou Hamdan, Ioannis D. Schizas, David L. Boothe, and Vasileios Maroulas. Bayesian sheaf neural networks. *arXiv preprint arXiv:2410.09590*, 2024. Revised September 2025.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Kyle Helfrich, Devin Willmott, and Qiang Ye. Orthogonal recurrent neural networks with scaled Cayley transform. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1969–1978. PMLR, 2018.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *Proceedings of the 7th International Conference on Learning Representations*, 2019.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *Proceedings of the 5th International Conference on Learning Representations*, 2017.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

Lei Huang, Xianglong Liu, Bo Lang, Adams Wei Yu, Yongliang Wang, and Bo Li. Orthogonal weight normalization: Solution to optimization over multiple dependent Stiefel manifolds in deep neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1):3271–3278, 2018. doi: 10.1609/aaai.v32i1.11768.

Saurav Jha, Dong Gong, and Lina Yao. CLAP4CLIP: Continual learning with probabilistic finetuning for vision-language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-4102.

Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999. doi: 10.1023/A:1007665907178.

Mohammad Emtiyaz Khan and Håvard Rue. The Bayesian learning rule. *Journal of Machine Learning Research*, 24(281):1–46, 2023.

Mohammad Emtiyaz Khan, Didrik Nielsen, Voot Tangkaratt, Wu Lin, Yarin Gal, and Akash Srivastava. Fast and scalable Bayesian deep learning by weight-perturbation in Adam. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2611–2620. PMLR, 2018.

Diederik P. Kingma, Tim Salimans, and Max Welling. Variational dropout and the local reparameterization trick. In *Advances in Neural Information Processing Systems*, volume 28, pp. 2575–2583. Curran Associates, Inc., 2015.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

David Krueger, Chin-Wei Huang, Riashat Islam, Ryan Turner, Alexandre Lacoste, and Aaron Courville. Bayesian Hypernetworks. *arXiv preprint arXiv:1710.04759*, 2017.

Mario Lezcano-Casado. Trivializations for gradient-based optimization on manifolds. In *Advances in Neural Information Processing Systems*, volume 32, pp. 9154–9164. Curran Associates, Inc., 2019.

Jianhua Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991. doi: 10.1109/18.61115.

Christos Louizos and Max Welling. Structured and efficient variational deep learning with matrix gaussian posteriors. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1708–1716. PMLR, 2016.

Christos Louizos and Max Welling. Multiplicative normalizing flows for variational Bayesian neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2218–2227. PMLR, 2017.

Wesley J. Maddox, Pavel Izmailov, Timur Garipov, Dmitry P. Vetrov, and Andrew Gordon Wilson. A simple baseline for Bayesian uncertainty in deep learning. In *Advances in Neural Information Processing Systems*, volume 32, pp. 13153–13164. Curran Associates, Inc., 2019.

Stephan Mandt, Matthew D. Hoffman, and David M. Blei. Stochastic gradient descent as approximate Bayesian inference. *Journal of Machine Learning Research*, 18(134):1–35, 2017.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *Proceedings of the 6th International Conference on Learning Representations*, 2018.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NeurIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Sixth Indian Conference on Computer Vision, Graphics and Image Processing*, pp. 722–729, 2008. doi: 10.1109/ICVGIP.2008.47.

Kazuki Osawa, Siddharth Swaroop, Mohammad Emtiyaz Khan, Anirudh Jain, Runa Eschenhagen, Richard E. Turner, and Rio Yokota. Practical deep learning with Bayesian principles. In *Advances in Neural Information Processing Systems*, volume 32, pp. 4287–4299. Curran Associates, Inc., 2019.

Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1530–1538. PMLR, 2015.

Yuesong Shen, Nico Daheim, Bai Cong, Peter Nickl, Gian Maria Marconi, Clement Bazan, Rio Yokota, Iryna Gurevych, Daniel Cremers, Mohammad Emtiyaz Khan, and Thomas Möllenhoff. Variational learning is effective for large deep networks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 44665–44686. PMLR, 2024.

Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 681–688. Omnipress, 2011.

Yeming Wen, Paul Vicol, Jimmy Ba, Dustin Tran, and Roger Grosse. Flipout: Efficient pseudo-independent weight perturbations on mini-batches. In *Proceedings of the 6th International Conference on Learning Representations*, 2018.

Yeming Wen, Dustin Tran, and Jimmy Ba. BatchEnsemble: An alternative approach to efficient ensemble and lifelong learning. In *International Conference on Learning Representations*, 2020.

A Appendix

A.1 Proofs

A.1.1 Proof of Theorem 1

Fix the nonsampled reference matrix W° as in Theorem 1. Write $f = f_{W^\circ}$ for the same measurable constructor as in equation 2. Define the latent likelihood and the Radon–Nikodym derivative by

$$L_{W^\circ}(R) = p(\mathcal{D} \mid f(R)), \qquad r(R) = \frac{dq_R}{dp_R}. \tag{18}$$

Here p_R is the latent prior, q_R is its variational posterior, and $q_R(\mathrm{d}R) = r(R)p_R(\mathrm{d}R)$. The likelihood L_{W° is nonnegative, and $q_R \ll p_R$ implies $r > 0$ q_R-almost surely. Consequently, restriction to $\{r > 0\}$ and the change of measure give

$$\begin{aligned} p(\mathcal{D} \mid W^\circ) &= \int L_{W^\circ}(R)\, p_R(\mathrm{d}R) \\ &\geq \int_{\{r>0\}} L_{W^\circ}(R)\, p_R(\mathrm{d}R) = \int \frac{L_{W^\circ}(R)}{r(R)}\, q_R(\mathrm{d}R). \end{aligned} \tag{19}$$

The conditional evidence $p(\mathcal{D} \mid W^\circ)$ is finite and positive by assumption. The assumed finite expected log-likelihood and finite latent KL make the following Jensen step well-defined:

$$\begin{aligned} \log p(\mathcal{D} \mid W^\circ) &\geq \log \mathbb{E}_{q_R}\left[\frac{L_{W^\circ}(R)}{r(R)}\right] \\ &\geq \mathbb{E}_{q_R}\left[\log L_{W^\circ}(R) - \log r(R)\right] \\ &= \mathbb{E}_{q_R}\left[\log p(\mathcal{D} \mid f_{W^\circ}(R))\right] - \mathrm{KL}(q_R \| p_R). \end{aligned} \tag{20}$$

This is equation 8. When the Gaussian prior and posterior in equation 4 are nondegenerate, $r > 0$ also holds p_R-almost surely. In that case the restriction in equation 19 is an equality.

For the second claim, let $p_W = f_\# p_R$ and $q_W = f_\# q_R$ be the induced weight distributions defined in the theorem. Measurability of f and $q_R \ll p_R$ imply $q_W \ll p_W$. Their Radon–Nikodym derivative satisfies

$$\frac{dq_W}{dp_W}(f(R)) = \mathbb{E}_{p_R}[r(R) \mid f(R)] \qquad p_R\text{-almost surely.} \tag{21}$$

Here conditioning on f means conditioning on the generated weight $f(R)$, or equivalently on the sigma algebra it generates. For any bounded measurable test function h on weight space, integration against $h(f(R))p_R(\mathrm{d}R)$ gives $\mathbb{E}_{q_R}[h(f(R))]$ on both sides. This identifies the density ratio on the image, regardless of whether p_W has a density with respect to ambient Lebesgue measure. With $\varphi(u) = u \log u$ for $u > 0$ and $\varphi(0) = 0$, conditional Jensen's inequality gives

$$\begin{aligned} \mathrm{KL}(q_W \| p_W) &= \mathbb{E}_{p_R}\left[\varphi(\mathbb{E}_{p_R}[r \mid f])\right] \\ &\leq \mathbb{E}_{p_R}\left[\mathbb{E}_{p_R}[\varphi(r) \mid f]\right] \\ &= \mathbb{E}_{p_R}[\varphi(r)] = \mathrm{KL}(q_R \| p_R). \end{aligned} \tag{22}$$

This proves $\mathrm{KL}(q_W \| p_W) \leq \mathrm{KL}(q_R \| p_R)$. Both inequalities hold for each fixed W°; optimizing W° in equation 6 varies the corresponding conditional evidence and its lower bound together. □

A.1.2 Proof of Theorem 2

Fix $W^\circ \in \mathbb{R}^{m\times d}$ and $R \in \mathbb{R}^{d\times d}$ as in Theorem 2, and abbreviate $A = A(R) = \frac{1}{2}(R - R^\top)$ and $Q = Q(R)$. Thus $A^\top = -A$. For every nonzero $x \in \mathbb{R}^d$,

$$x^\top (I + A)x = \|x\|_2^2 > 0, \tag{23}$$

because $x^\top Ax = 0$. Hence $I + A$ is invertible; the same argument applies to $I - A$. Since these two matrices commute,

$$Q^\top = (I - A)^{-1}(I + A), \tag{24}$$

$$Q^\top Q = (I - A)^{-1}(I + A)(I - A)(I + A)^{-1} = I. \tag{25}$$

Thus $Q \in \mathrm{O}(d)$. The continuous path

$$s \longmapsto (I - sA)(I + sA)^{-1}, \qquad s \in [0, 1], \tag{26}$$

lies in $\mathrm{O}(d)$ and starts at I. Its determinant is therefore constantly 1, proving $Q \in \mathrm{SO}(d)$.

The special orthogonal group is $\mathrm{SO}(d) = \{U \in \mathbb{R}^{d\times d} : U^\top U = I,\ \det U = 1\}$. Moreover,

$$I + Q = 2(I + A)^{-1} \tag{27}$$

is invertible, so $-1 \notin \mathrm{spec}(Q)$. Conversely, for $Q \in \mathrm{SO}(d)$ with $I + Q$ invertible, define

$$A = (I - Q)(I + Q)^{-1}. \tag{28}$$

Using $Q^\top Q = I$ gives $A^\top = -A$, and direct substitution recovers $Q = (I - A)(I + A)^{-1}$. Thus the Cayley image consists precisely of rotations without eigenvalue -1. Every element of $\mathrm{SO}(d)$ admits an orthogonal block decomposition into planar rotations and $+1$ blocks; its -1 eigenvalues occur in pairs and can be written as planar rotations by π. Replacing these angles by values tending to π produces rotations in the Cayley image converging to the original rotation. Therefore the image is dense in $\mathrm{SO}(d)$. Since the orbit $\mathcal{O}(W^\circ) = \{W^\circ U : U \in \mathrm{SO}(d)\}$ is compact and the orbit map is continuous,

$$\overline{\{f_{W^\circ}(R) : R \in \mathbb{R}^{d\times d}\}} = \mathcal{O}(W^\circ). \tag{29}$$

For the sampled weight $W = f_{W^\circ}(R) = W^\circ Q^\top$ from equation 2, orthogonality gives

$$WW^\top = W^\circ Q^\top Q W^{\circ\top} = W^\circ W^{\circ\top}. \tag{30}$$

Taking traces yields $\|W\|_F = \|W^\circ\|_F$, and multiplication by an orthogonal matrix preserves every singular value. This proves the invariance claims in Theorem 2.

We next compute the constructor differential at the latent center $M \in \mathbb{R}^{d\times d}$. Set $A_0 = A(M)$ and denote the transpose Cayley map by

$$\mathcal{C}(A) = (I - A)^{-1}(I + A), \qquad \mathcal{C}(A(R)) = Q(R)^\top. \tag{31}$$

Here $\mathfrak{so}(d) = \{C \in \mathbb{R}^{d\times d} : C^\top = -C\}$ is the space of skew-symmetric matrices. For a direction $C \in \mathfrak{so}(d)$, differentiation of the matrix inverse gives

$$\begin{aligned} \mathrm{D}\mathcal{C}(A)[C] &= (I - A)^{-1}C(I - A)^{-1}(I + A) + (I - A)^{-1}C \\ &= 2(I - A)^{-1}C(I - A)^{-1}. \end{aligned} \tag{32}$$

For a latent perturbation $H \in \mathbb{R}^{d\times d}$, the skew extraction is linear. Thus $\mathrm{D}A(M)[H] = A(H)$, where $A(H) = \frac{1}{2}(H - H^\top)$ as in the main text. The chain rule yields

$$\mathrm{D}f_{W^\circ}(M)[H] = 2W^\circ(I - A_0)^{-1}A(H)(I - A_0)^{-1} =: \mathcal{L}_M(H). \tag{33}$$

The notation $\mathcal{L}_M(H)$ in equation 33 denotes this linear map of H. Smoothness of the inverse on skew matrices gives, for each fixed H,

$$f_{W^\circ}(M + sH) = W_M + s\mathcal{L}_M(H) + O(s^2), \qquad s \to 0. \tag{34}$$

The orbit tangent space at $W_M = W^\circ\mathcal{C}(A_0)$ is

$$T_{W_M}\mathcal{O}(W^\circ) = \{W_M\Omega : \Omega \in \mathfrak{so}(d)\}. \tag{35}$$

Indeed, $s \mapsto W_M \exp(s\Omega)$, where exp is the matrix exponential, is an orbit curve with derivative $W_M\Omega$; every orbit tangent vector arises by differentiating the group action. Since $\mathcal{C}(A)$ is orthogonal, differentiation of $\mathcal{C}(A)^\top\mathcal{C}(A) = I$ shows that $\Omega_H := \mathcal{C}(A_0)^\top\mathrm{D}\mathcal{C}(A_0)[A(H)]$ is skew-symmetric. Hence

$$\mathrm{D}f_{W^\circ}(M)[H] = W_M\Omega_H \tag{36}$$

so $\mathrm{Im}\,\mathcal{L}_M \subseteq T_{W_M}\mathcal{O}(W^\circ)$. Conversely, $H \mapsto A(H)$ is onto $\mathfrak{so}(d)$: take $H = C$ for any skew C. Thus equation 32 is injective on this $d(d-1)/2$ dimensional space, because $(I - A_0)^{-1}$ is invertible. Its image is therefore the entire tangent space of $\mathrm{SO}(d)$ at $\mathcal{C}(A_0)$, which has the same dimension.

Multiplying by W° maps that group tangent space onto equation 35, even when the orbit map itself has a kernel because W° is rank deficient. Consequently, $\operatorname{Im} \mathcal{L}_M = T_{W_M}\mathcal{O}(W^\circ)$.

Finally, if H is symmetric, then $A(R+H) = A(R)$ for every R: these latent directions are globally removed by the constructor. At a center with $A(M) = 0$, equation 33 simplifies to

$$\mathcal{L}_M(H) = 2W^\circ A(H). \tag{37}$$

Here $H = H_{\mathrm{s}} + \Omega$ with $H_{\mathrm{s}} = \frac{1}{2}(H + H^\top)$ symmetric and $\Omega = A(H)$ skew-symmetric. Thus $\mathcal{L}_M(H) = 0$ exactly when $W^\circ\Omega = 0$ at such a center. This also identifies any additional null directions arising from a nontrivial stabilizer of the reference matrix. □

A.1.3 Proof of Theorem 3

We compare the two random weights defined in Theorem 3 around the same center $W_M = f_{W^\circ}(M)$. Write $\mathcal{O} = \mathcal{O}(W^\circ)$ and identify each $m \times d$ matrix with its vectorization in $\mathbb{R}^{md}$. At W_M, P_T is the orthogonal projector onto $\operatorname{vec}(T_{W_M}\mathcal{O})$ from equation 35, and $P_N = I_{md} - P_T$ projects onto its orthogonal complement. The orbit $\mathcal{O}$ is the continuous image of compact $\mathrm{SO}(d)$, so it is compact and closed. It is a smooth embedded submanifold because it is an orbit of a smooth compact-group action.

Because the orbit is lower-dimensional by assumption, it has zero ambient Lebesgue measure. For $t > 0$, $\operatorname{vec}(W_{\mathrm{E}}(t)) = w_M + t\xi$ has covariance $t^2 D \succ 0$ and hence an ambient density. Thus $W_{\mathrm{E}}(t) \notin \mathcal{O}$ almost surely. Since $\mathcal{O}$ is closed, its distance $d_{\mathcal{O}}(W_{\mathrm{E}}(t))$ is positive almost surely. Also, using $W_M \in \mathcal{O}$ shows $d_{\mathcal{O}}(W_{\mathrm{E}}(t)) \le t\|\xi\|_2$, so its second moment is finite.

To compute the leading expected distance, use a tubular neighborhood of the compact smooth orbit. There the nearest-point projection is smooth, so squared distance to $\mathcal{O}$ has value and gradient zero at W_M and Hessian $2P_N$ in vectorized coordinates. Smoothness through third order in this neighborhood gives constants $c, \eta > 0$,

$$\left|d_{\mathcal{O}}(W_M + h)^2 - \|P_N \operatorname{vec}(h)\|_2^2\right| \le c\|h\|_F^3 \quad \text{whenever } \|h\|_F \le \eta. \tag{38}$$

For $t > 0$, set $E_t = \{t\|\xi\|_2 \le \eta\}$. On E_t, apply equation 38 to $h = \operatorname{mat}_{m\times d}(t\xi)$; the expected absolute remainder is at most $ct^3\mathbb{E}\|\xi\|_2^3$. On E_t^c,

$$d_{\mathcal{O}}(W_{\mathrm{E}}(t)) \le t\|\xi\|, \qquad \|P_N t\xi\| \le t\|\xi\|, \tag{39}$$

because $W_M \in \mathcal{O}$ and P_N is an orthogonal projector. Both expected squared terms on E_t^c are bounded by

$$t^2\mathbb{E}\left[\|\xi\|_2^2 \mathbf{1}_{E_t^c}\right] \le \frac{t^4}{\eta^2}\mathbb{E}\|\xi\|_2^4 = O(t^4), \tag{40}$$

using $\mathbf{1}_{E_t^c} \le t^2\|\xi\|_2^2/\eta^2$ and the finite Gaussian fourth moment. Therefore

$$\mathbb{E}[d_{\mathcal{O}}(W_{\mathrm{E}}(t))^2] = t^2\mathbb{E}\|P_N\xi\|^2 + O(t^3) = t^2 \operatorname{tr}(P_N D) + O(t^3). \tag{41}$$

Here $\mathbb{E}\|P_N\xi\|_2^2 = \operatorname{tr}(P_N D)$ because $P_N^2 = P_N$ and $\operatorname{Cov}(\xi) = D$. Since $P_N \neq 0$ and $D \succ 0$, this trace is strictly positive. By the same trace identity, the additive Gaussian's tangent energy is exactly

$$\mathbb{E}\|P_T \operatorname{vec}(W_{\mathrm{E}}(t) - W_M)\|_2^2 = t^2 \operatorname{tr}(P_T D). \tag{42}$$

For the UniCOP draw $W_{\mathrm{G}}(s) = f_{W^\circ}(M + sZ)$, let $C = A(Z)$ and $A_0 = A(M)$ be the skew components of Z and M, respectively. Along $A_0 + sC$, write $B_s = (I - A_0 - sC)^{-1}$. By equation 32, the first two derivatives of $\mathcal{C}(A_0 + sC)$ are $2B_sCB_s$ and $4B_sCB_sCB_s$. If $K^\top = -K$, then $\|(I-K)x\|_2^2 = \|x\|_2^2 + \|Kx\|_2^2$, so

$$\|(I-K)^{-1}\|_2 \le 1. \tag{43}$$

Here $\|W^\circ\|_2$ denotes the spectral operator norm. In particular $\|B_s\|_2 \le 1$ for every real s, and the second derivative of $f_{W^\circ}(M + sZ)$ has Frobenius norm at most $4\|W^\circ\|_2\|C\|_F^2$. Taylor's theorem with integral remainder therefore gives

$$\begin{aligned} f_{W^\circ}(M + sZ) &= W_M + s\mathcal{L}_M(Z) + \mathcal{R}_s(Z), \\ \|\mathcal{R}_s(Z)\|_F &\le 2\|W^\circ\|_2 s^2 \|A(Z)\|_F^2 \end{aligned} \tag{44}$$

uniformly in $s \in \mathbb{R}$ and Z. Also, equation 33 and $\|B_0\|_2 \le 1$ give $\|\mathcal{L}_M(Z)\|_F \le 2\|W^\circ\|_2\|A(Z)\|_F$. The fourth-moment assumption on Z therefore controls both $\mathcal{L}_M(Z)$ and $\mathcal{R}_s(Z)$. Since $\mathcal{L}_M(Z) \in T_{W_M}\mathcal{O}$ by Theorem 2, $P_T \operatorname{vec}(\mathcal{L}_M(Z)) = \operatorname{vec}(\mathcal{L}_M(Z))$. Projecting equation 44 and expanding its squared norm yields

$$\mathbb{E}\|P_T \operatorname{vec}(W_{\mathrm{G}}(s) - W_M)\|_2^2 = s^2\mathbb{E}\|\mathcal{L}_M(Z)\|_F^2 + O(s^3). \tag{45}$$

The linear–remainder cross term is $O(|s|^3\mathbb{E}\|Z\|_F^3)$ and the remainder-square term is $O(s^4\mathbb{E}\|Z\|_F^4)$. Both are finite under the stated fourth-moment assumption. For every real s and every realization of Z, equation 2 lies in $\mathcal{O}$, so exact orbit support gives $d_{\mathcal{O}}(W_{\mathrm{G}}(s)) = 0$ almost surely.

Finally, $c_{\mathrm{E}} = \operatorname{tr}(P_T D)$ and $c_{\mathrm{G}} = \mathbb{E}\|\mathcal{L}_M(Z)\|_F^2$ are the strictly positive tangent-energy coefficients defined in the theorem. Taking $s = s(t) = t\sqrt{c_{\mathrm{E}}/c_{\mathrm{G}}}$ in equation 45 gives

$$\mathbb{E}\|P_T \operatorname{vec}(W_{\mathrm{G}}(s) - W_M)\|_2^2 = t^2 c_{\mathrm{E}} + O(t^3), \tag{46}$$

which matches the additive tangent energy in equation 42 to leading order while its off-orbit leakage remains exactly zero. □

A.1.4 Proof of Theorem 4

Fix the reference W° and the model inputs X in Theorem 4. The map $g_X(R)$ returns the vector of model outputs obtained using the sampled weight $f_{W^\circ}(R)$. To use the vector Gaussian distributions in the theorem, set $n = n_R = d^2$, $m = \operatorname{vec}(M)$, and define $h : \mathbb{R}^n \to \mathbb{R}^k$ by $h(r) = g_X(\operatorname{mat}_{d\times d}(r))$, where k is the number of output coordinates on X. Then $h(m) = g_X(M)$, and its $k \times n$ Jacobian at m is $J = \mathrm{D}h(m) = J_X$. For $\xi \sim \mathcal{N}(0, D)$, the random vector $r_t = m + t\xi$ has law q_t, so $h(r_t) = g_X(R_t)$. The bounded second derivative of g_X (equivalently of h under this fixed reshaping) yields a constant $L < \infty$ such that Taylor's theorem gives

$$h(m + t\xi) = h(m) + tJ\xi + \rho_t(\xi), \qquad \|\rho_t(\xi)\| \le \frac{L}{2}t^2\|\xi\|^2 \tag{47}$$

Here $\rho_t(\xi)$ is the Taylor remainder, and $\|\cdot\|$ on output vectors is the Euclidean norm. Since a fixed-dimensional Gaussian has finite moments of every order,

$$\mathbb{E}\|\rho_t\|^2 = O(t^4), \qquad \mathbb{E}\|J\xi\|^2 = O(1). \tag{48}$$

Because $\mathbb{E}\xi = 0$, centering equation 47 gives $h(m + t\xi) - \mathbb{E}h(m + t\xi) = tJ\xi + \rho_t(\xi) - \mathbb{E}\rho_t(\xi)$. Therefore

$$\operatorname{Cov}[h(m + t\xi)] = t^2 JDJ^\top + t\mathbb{E}\big[J\xi(\rho_t - \mathbb{E}\rho_t)^\top + (\rho_t - \mathbb{E}\rho_t)(J\xi)^\top\big] + \operatorname{Cov}(\rho_t). \tag{49}$$

By Cauchy–Schwarz, the matrix norm of the cross term is $O(t^3)$, while that of $\operatorname{Cov}(\rho_t)$ is $O(t^4)$. The same argument applies to $\widetilde{\xi} \sim \mathcal{N}(0, \widetilde{D})$, for which $\widetilde{r}_t = m + t\widetilde{\xi}$ has law $\widetilde{q}_t$. Thus, with the remainder measured in any fixed matrix norm,

$$\begin{aligned} \operatorname{Cov}[g_X(R_t)] &= t^2 J_X D J_X^\top + O(t^3), \\ \operatorname{Cov}[g_X(\widetilde{R}_t)] &= t^2 J_X \widetilde{D} J_X^\top + O(t^3). \end{aligned} \tag{50}$$

Next, the isotropic latent prior in equation 4 is $p_R = \mathcal{N}(0, \sigma_0^2 I_n)$ when the full matrix is vectorized. The Gaussian KL formula gives, for $t > 0$,

$$\mathrm{KL}(q_t \| p_R) = \frac{1}{2}\left[\frac{\|m\|_2^2 + t^2 \operatorname{tr}(D)}{\sigma_0^2} - n + n\log\sigma_0^2 - \log\det(t^2 D)\right]. \tag{51}$$

The trace term measures total latent variance; the determinant term measures the product of its eigenvalues. Since the permutation matrix Π is orthogonal, $\widetilde{D} = \Pi D \Pi^\top$ has exactly the same trace and determinant as D, while the common mean is m. Thus replacing D with $\widetilde{D}$ leaves equation 51 unchanged for every $t > 0$. Subtracting the two covariance expansions in equation 50 gives

$$\operatorname{Cov}[g_X(R_t)] - \operatorname{Cov}[g_X(\widetilde{R}_t)] = t^2 J(D - \widetilde{D})J^\top + O(t^3), \tag{52}$$

which proves the second claim of Theorem 4. To interpret the dependence on orientation in latent space, write $D = \mathrm{diag}(d_1, \ldots, d_n)$ and let J_ℓ denote the ℓth column of J. Then

$$JDJ^\top = \sum_{\ell=1}^{n} d_\ell J_\ell J_\ell^\top. \tag{53}$$

For example, if the permutation only swaps entries d_i and d_j, subtraction gives

$$J(D - \widetilde{D})J^\top = (d_i - d_j)(J_i J_i^\top - J_j J_j^\top). \tag{54}$$

Thus equal latent KL produces different leading output covariance precisely when the resulting matrix $J(D - \widetilde{D})J^\top$ is nonzero; equal KL alone does not guarantee a change.

This interpretation is consistent with the constructor's null directions. If an R-shaped random perturbation H is symmetric almost surely, then $A(M + H) = A(M)$ almost surely. By equation 2,

$$Q(M + H) = Q(M), \qquad f_{W^\circ}(M + H) = f_{W^\circ}(M), \qquad g_X(M + H) = g_X(M) \tag{55}$$

almost surely, so the induced weight and function covariances are zero. □

A.2 Experimental Settings

A.2.1 Computing Resources

We conducted the experiments on a server equipped with eight NVIDIA GeForce RTX 4090 GPUs with 192 CPU cores and 256G memory. The dataset, backbone, training schedule, posterior-sample count, and number of independent runs for each study are specified in the corresponding experimental sections below.

A.2.2 Baselines

We compare UniCOP with the following predictive models, posterior inference mechanisms, and task-specific controls. Unless stated otherwise, a "+UniCOP" variant retains the corresponding inference mechanism and replaces its weight-sampling construction with the Cayley-orbit constructor.

- **Deterministic DNN** (He et al., 2016): a point-estimate network, used as a non-Bayesian reference for in-domain classification.
- **BatchEnsemble** (Wen et al., 2020): an ensemble that shares a base weight matrix across members and uses member-specific rank-one factors.
- **MC Dropout** (Gal & Ghahramani, 2016): dropout remains active at test time, and predictions are averaged across stochastic forward passes.
- **SWAG** (Maddox et al., 2019): an approximate Gaussian weight posterior constructed from stochastic-optimizer trajectories.
- **Variational Dropout (VD)** (Kingma et al., 2015): a variational weight-uncertainty baseline using learned dropout rates.
- **Radial BNN** (Farquhar et al., 2020): a radial variational posterior used as a weight-space uncertainty baseline.
- **Bayes by Backprop (BBB)** (Blundell et al., 2015): an explicit variational mechanism that learns the mean and scale of a diagonal-Gaussian weight posterior. BBB+UniCOP instead samples latent variables and maps them to weights using the shared constructor.
- **Rank-1 BNN** (Dusenberry et al., 2020): a variational mechanism that samples low-dimensional rank-one factors. Its UniCOP pairing tests the constructor beyond elementwise BBB sampling.
- **VOGN** (Osawa et al., 2019): a variational online Gauss–Newton mechanism that maintains approximate posterior statistics during optimization; VOGN+UniCOP uses the same posterior constructor as the other pairings.
- **IVON** (Shen et al., 2024): an optimizer-induced mechanism that maintains posterior precision statistics. IVON+UniCOP uses those statistics to determine latent sampling scales before applying the Cayley map.

- **Matched Euclidean decoder**: in the fixed-posterior IVON control, a center-matched, locally energy-matched additive map decodes the *same* latent posterior and noise draws as UniCOP. Its construction is given in Appendix A.3.2.
- **CLAP4CLIP Gaussian feature posterior** (Jha et al., 2024): a diagonal-Gaussian residual over text features, compared with blockwise Cayley rotations while retaining the same downstream backbone and training protocol (Appendix A.4).
- **Orbit-aligned reference posterior**: in the controlled two-dimensional task, numerical integration over a dense angle grid gives a reference posterior and predictive distribution. BBB, IVON, and their UniCOP variants are evaluated against this common target (Appendix A.3.1).

A.2.3 Evaluation Metrics

For classification, let $\bar{p}_i(c)$ be the posterior-predictive probability of class c for example i, obtained by averaging predictive probabilities over weight samples; y_i is its true class and $\hat{y}_i = \arg\max_c \bar{p}_i(c)$. We use the following metrics, with the preferred direction indicated in the tables and figures.

- **Top-1 accuracy (↑)**: the proportion of examples for which $\hat{y}_i = y_i$. Classification accuracy is reported in percent.
- **Negative log likelihood (NLL, ↓)**: $-N^{-1}\sum_{i=1}^{N}\log\bar{p}_i(y_i)$, measuring the probability assigned to the true class.
- **Expected calibration error (ECE, ↓)**: the bin-frequency-weighted mean absolute gap between empirical accuracy and mean confidence across confidence bins (Guo et al., 2017). ECE is reported in percent in the main tables.
- **Brier score (↓)**: $N^{-1}\sum_{i=1}^{N}\sum_{c}\big(\bar{p}_i(c) - \mathbf{1}\{y_i = c\}\big)^2$. Tables headed Brier(%) multiply this value by 100.
- **OOD AUROC (↑) and AUPR-out (↑)**: the areas under the receiver-operating-characteristic and precision–recall curves, respectively. OOD examples are the positive class for AUPR-out. In discrete OOD detection, the confidence score is $\max_c \bar{p}_i(c)$ and lower confidence indicates OOD; the CLAP4CLIP future-class experiment instead uses the selected predictive-MI score (Appendix A.4).
- **FPR95 (↓)**: the OOD false-positive rate at the first threshold reaching at least 95% in-domain true-positive rate. **Detection error (↓)** is the minimum, over decision thresholds, of the mean of the in-domain false-negative and OOD false-positive rates.
- **Selective risk, AURC, and E-AURC (↓)**: selective risk is the error rate among retained predictions after ordering examples by confidence or uncertainty; AURC integrates risk across coverage levels, and E-AURC subtracts the corresponding oracle risk–coverage area. In the corruption experiments, posterior-predictive entropy ranks predictions for selective evaluation.
- **Predictive mutual information (MI)** (Houlsby et al., 2011): $H(\bar{p}_i) - S^{-1}\sum_{s=1}^{S} H(p_i^{(s)})$, separating variation among S posterior predictive samples from their mean predictive entropy. **Pairwise Jensen–Shannon divergence (JSD)** (Lin, 1991) averages the JSD between distinct sample-predictive distributions. In the fixed-posterior control, both scores rank errors of the shared center predictor; their AUROC is higher and AURC lower when this ranking is better.
- **Relative row-Gram drift (↓)**: the relative Frobenius change in $WW^\top$ from the reference $W^\circ W^{\circ\top}$, measuring violation of the structure preserved by UniCOP. The CLAP4CLIP geometry diagnostic analogously measures relative squared-norm drift within text-feature blocks; median angular displacement measures directional variation.
- **Reference-posterior discrepancies (↓)**: predictive KL averages the Bernoulli KL from the grid-reference predictive to a model predictive over test directions; angular W_2 compares posterior angle distributions; weight energy distance compares sampled weight distributions; and dense NLL is cross-entropy against the reference predictive on the same directions (Appendix A.3.1).

- **Class-incremental accuracy and backward transfer** (↑): CLAP4CLIP reports final average accuracy over learned classes (Acc. avg), accuracy on the final task (Acc. last), and backward transfer (BWT), which summarizes how learning subsequent sessions changes performance on earlier tasks. Future-class OOD AUROC, AUPR, and FPR95 are averaged over sessions 0–8 within each held-out class order before aggregation across orders.

A.3 Mechanistic Evidence Across Posterior Inference Mechanisms

A.3.1 Orbit-aligned posterior under explicit variational inference

For each of 20 paired problems, we fixed a two-dimensional reference weight $w^\circ = (2, 0)$ and generated 28 binary observations at input directions spanning the unit circle, with small direction jitter. Labels came from a rotated copy of w°; the rotation and observations varied across problems. The reference model restricts the weight to $w(\theta) = R(\theta)w^\circ$ and places a $\mathcal{N}(0, 0.35^2)$ prior on θ. We evaluated its normalized likelihood times prior on 4,097 angles in $[-1.35, 1.35]$, then integrated the resulting posterior to obtain a reference predictive. This supplies a known curved target against which both weight-space and predictive approximations can be measured.

We fitted four models to the same observations in each problem: ambient Gaussian BBB and IVON posteriors, and their UniCOP counterparts. BBB learns a mean-field Gaussian directly over the two weights; BBB+UniCOP learns a Gaussian over latent angle parameters and pushes its samples through the Cayley map:

$$W_{\mathrm{BBB}} = \mu + \sigma \odot \epsilon, \qquad W_{\mathrm{BBB+UniCOP}} = f_{W^\circ}(M + \exp(S) \odot \epsilon)\,. \tag{56}$$

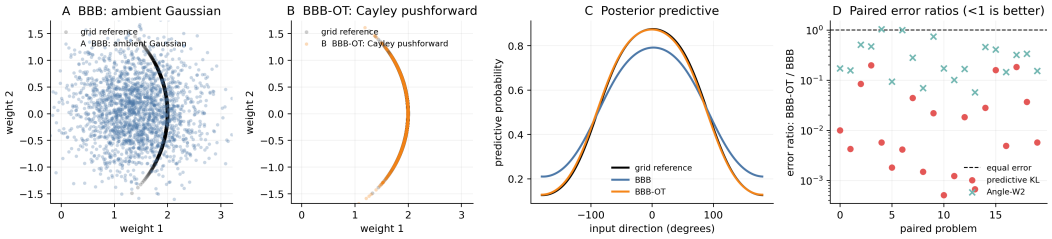


Figure 4: **Curved posterior under explicit variational inference.** BBB-OT denotes BBB+UniCOP. Panels A–C show posterior weight samples and predictive probabilities against the grid reference for a representative problem. Panel D shows, for all 20 problems, the BBB+UniCOP-to-BBB ratios of predictive KL and angular W_2; values below one favor UniCOP.

The UniCOP reference weight was held fixed so its samples lie on the target radius. BBB used a weight prior with standard deviation 2.0; BBB+UniCOP used a latent prior with standard deviation 0.25. Both optimized the Monte Carlo estimate of mean binary cross-entropy plus the respective layer KL divided by 28, using 1,000 full-batch Adam steps at learning rate 0.025 and four posterior samples per step. The IVON rows used the same problems, steps, and sample count, with learning rate 0.035 and effective sample size 28; their posterior scales came from optimizer precision rather than learned BBB log-scales. Consequently, the four rows compare complete inference constructions: parameterization, prior, and KL space differ, so their gaps alone do not identify a matched-KL causal effect.

For evaluation, we drew 2,048 weight samples per fitted model and sampled the grid posterior at corresponding quantiles. Predictive KL is the average Bernoulli $D_{\mathrm{KL}}(p_{\mathrm{ref}} \| p_{\mathrm{model}})$ over 181 test directions. Angular W_2 compares empirical orientations with reference angles; weight energy distance compares the two weight-sample distributions. Dense NLL is cross-entropy against the reference predictive on the same directions, and therefore includes the reference predictive entropy. All four metrics in Table 6 are lower when the approximation is closer to the reference.

A.3.2 Fixed-posterior control under optimizer-induced inference

IVON maintains precision statistics during training and uses them to determine latent posterior scales. With latent mean $M_t = \mathrm{vec}(\texttt{angle}_t)$, precision π_t, prior contribution $\lambda_A = (N_{\mathrm{eff}}\sigma_0^2)^{-1}$,

Table 6: **Curved-posterior task.** Means over 20 paired problems; lower is better for all metrics. The IVON rows from the same factorial experiment show that the curved-target trend also occurs with optimizer-induced inference.

Method	Predictive KL ↓	Angle-W_2 ↓	Weight energy ↓	Dense NLL ↓
BBB	0.015066	0.230928	1.131163	0.536515
BBB+**UniCOP**	0.000218	0.051800	0.488771	0.521667
IVON	0.015848	0.242085	1.151585	0.537297
IVON+**UniCOP**	0.001087	0.114522	0.626816	0.522535

and Gaussian noise ϵ, IVON+UniCOP uses

$$S_t = -\tfrac{1}{2}\log[N_{\text{eff}}(\pi_t + \lambda_A)]\,, \qquad z_t = M_t + \exp(S_t) \odot \epsilon, \tag{57}$$

$$W_t = f_{W_t^\circ}(\text{mat}(z_t)) = W_t^\circ Q(\text{mat}(z_t))^\top. \tag{58}$$

Only latent parameters are sampled; reference weights, biases, and optional scales remain point estimates. The control decodes the same latent posterior through a center-matched, locally energy-matched Euclidean map:

$$T_{\text{E}}(Z) = I + \gamma_d Z, \qquad \gamma_d = \sqrt{\frac{2(d-1)}{d}}, \tag{59}$$

$$W_{\text{E}}^{\text{ctrl}}(Z) = W^\circ Q(M)^\top T_{\text{E}}(Z - M)^\top. \tag{60}$$

Thus the two decoders have identical center predictions, latent posterior states, noise draws, and latent KL; only their maps from latent variables to weights differ.

We evaluated five IVON+UniCOP CIFAR-10/ResNet-20 models (Krizhevsky, 2009; He et al., 2016) trained for 200 epochs across seven uncertainty budgets, using 64 posterior samples at each seed–budget point on the official test split. The budgets scale the covariance of $q_\alpha(z) = \mathcal{N}(\mu, \alpha^2\,\text{Diag}(\sigma^2))$ and are reported as $\kappa = \text{KL}(q_\alpha \| p)/d_z$. We measured relative row-Gram drift and ranked errors of the shared deterministic center predictor by predictive mutual information (MI) and pairwise Jensen–Shannon divergence (JSD). AUROC is higher when errors rank more accurately; AURC is lower when selective prediction is safer. The five training seeds, with budgets averaged within each seed, are the independent units for confidence intervals.

Table 7: **Matched IVON posterior by latent KL budget.** Mean effects over five seeds are UniCOP minus the center-matched Euclidean control. Euclidean drift is shown directly; UniCOP drift stays near 1.6×10^{-7} at every budget.

κ	Euclidean drift	ΔMI-AUROC	ΔMI-AURC	ΔJSD-AUROC
0.0046	2.156	+0.00884	−0.001188	+0.00335
0.0112	1.931	+0.00858	−0.001215	+0.00397
0.0271	1.684	+0.00657	−0.000898	+0.00315
0.0655	1.360	+0.00472	−0.000671	+0.00209
0.1586	0.969	+0.00268	−0.000496	+0.00140
0.3842	0.568	+0.00088	−0.000125	+0.00003
0.9304	0.213	+0.00107	−0.000229	+0.00072

Across budgets, UniCOP has mean row-Gram drift 1.58×10^{-7} versus 1.26883 for the Euclidean decoder. Mean MI-AUROC rises by 0.004764 (95% seed-bootstrap CI $[0.004029, 0.005539]$), MI-AURC falls by 0.000689 ($[-0.000781, -0.000605]$), and JSD-AUROC rises by 0.002101 ($[0.001736, 0.002630]$). Row-Gram preservation and higher MI-AUROC occur jointly at 34 of 35 seed–budget points. These results isolate a decoder-level effect under a fixed posterior, consistent with the structure-preserving and fixed-KL allocation analysis.

A.3.3 Additional Posterior-Geometry Diagnostics

We performed three further controlled diagnostics with the production BBB+UniCOP and IVON+UniCOP constructors. First, after matching their latent state and random draws, we com-

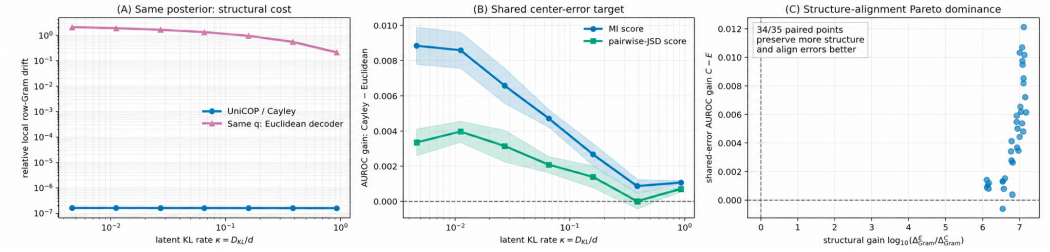


Figure 5: **IVON posterior-construction control.** UniCOP/Cayley in the archived plot denotes UniCOP. The two decoders share the learned IVON posterior, center prediction, and latent KL. UniCOP nearly eliminates row-Gram drift and improves uncertainty-based ranking of the shared center's errors at 34 of 35 seed–budget points.

pared their current weight samples and predictions while varying posterior precision and optimizer curvature separately. The two inference routes produced identical samples and predictions up to 4.44×10^{-16}; posterior precision changed current spread, whereas optimizer curvature changed the subsequent update. This checks that both mechanisms use the same constructor without conflating current posterior uncertainty with an optimizer step.

Second, we compared 4,096 Cayley particles with an ambient Gaussian control matched in root-mean-square displacement. The Cayley particles preserved the reference norm to relative error 3.69×10^{-16} and had radial standard deviation 1.14×10^{-16}, whereas the matched Gaussian had radial standard deviation 0.1875. The measured normal residual scaled approximately quadratically with latent perturbation size (log–log slope 1.999). These observations complement the orbit-support result of Theorem 2.

Third, an analytically specified Gaussian latent target allowed explicit BBB-style variational learning and IVON-induced learning to be assessed against the same posterior. After 900 optimization steps, the latent-mean root-mean-square errors were below 10^{-16} and 0.00119 for BBB and IVON, respectively; the respective energy distances from the exact Cayley pushforward were 1.38×10^{-9} and 0.000443. The trajectories differed, but both inference mechanisms recovered the common target posterior.

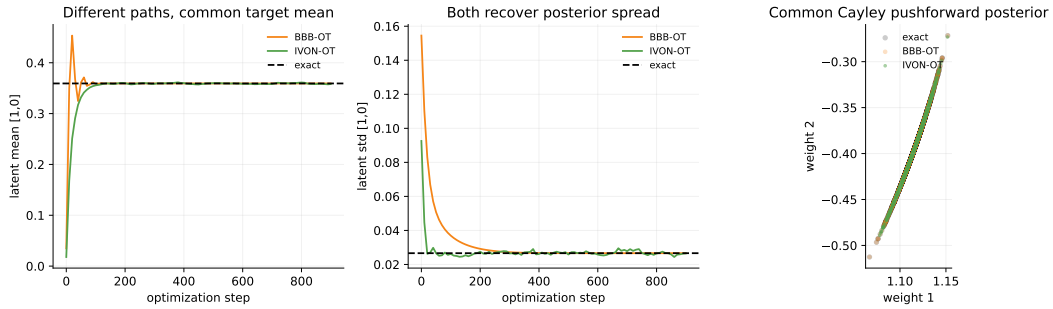


Figure 6: **One target, two inference mechanisms.** BBB+UniCOP and IVON+UniCOP approach the same analytic latent posterior through different optimization trajectories; the induced weight samples agree with the exact curved target.

We also tested the function-space effect of latent variance allocation under IVON+UniCOP. Across 20 problems, we kept the latent center, variance spectrum, and isotropic-prior KL fixed, then aligned or permuted variances among Lie directions; an equal-trace isotropic posterior served as a second control. Each setting used 4,096 posterior samples. Permutation changed output covariance despite zero KL difference: relative function-correlation error was 0.539 for the shuffled allocation and 0.358 for the isotropic control, relative to the aligned allocation. The corresponding pointwise W_2 distances in a target input gap were 0.01376 and 0.00558. This is a controlled demonstration of the allocation effect in Theorem 4, rather than an end-to-end training comparison.

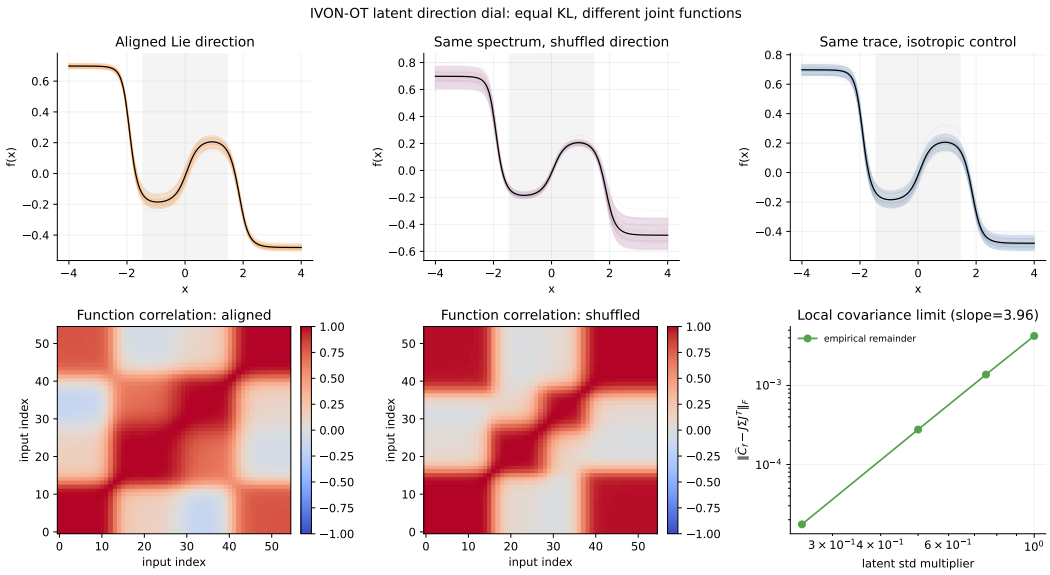


Figure 7: **Fixed-KL function-space allocation.** Aligning or permuting the same latent variance spectrum changes predictive uncertainty and joint function correlation while preserving latent KL. The lower-right panel checks the local covariance approximation.

A.4 Feature-Space Transfer in CLAP4CLIP

CLAP4CLIP (Jha et al., 2024) provides a downstream test of whether the posterior constructor transfers from network weights to text features. We used CLIP ViT-B/16 on CIFAR-100 class-incremental learning with ten sessions of ten newly introduced classes each. The baseline samples a diagonal-Gaussian residual around a 512-dimensional text feature. UniCOP divides that feature into 32 blocks of dimension 16 and applies a Cayley rotation within each block. The construction uses 3,840 latent angle coordinates and preserves each block's squared norm. Its latent prior and initial standard deviation are 0.1; the KL coefficient is dimension-normalized from 10^{-3} to $10^{-3}(512/3840)$.

Both systems retained the same backbone, replay, LASP and VGA components, optimizer settings, and training schedule: seven epochs in sessions 0–2, five in sessions 3–9, and two finetuning epochs. Predictions used 20 posterior samples. The posterior configuration and negative predictive mutual information (MI) OOD score were selected on development class orders and frozen before evaluation on held-out orders 0–2. Future classes are OOD at each session; OOD metrics average sessions 0–8 within each order before aggregation across orders. This is a posterior-system comparison: the feature-posterior parameter counts are not matched.

Table 8: **CLAP4CLIP held-out class orders.** Paired UniCOP effects in percentage points. Positive MI AUROC/AUPR and FPR95 improvement favor UniCOP; negative final average-accuracy differences show the accompanying ID cost.

Order	ΔMI AUROC	ΔMI AUPR	FPR95 improvement	Δfinal accuracy
0	+1.81	+6.81	+1.88	−0.34
1	+1.25	+5.79	+2.79	−0.55
2	+2.11	+6.25	+2.40	−0.73
Mean	+1.73	+6.28	+2.36	−0.54

The frozen MI score improves AUROC, AUPR, and FPR95 in all three held-out orders (Table 8). A mean-logit-energy sensitivity check also improves AUROC by 1.01 and AUPR by 1.92 points on average, while its FPR95 decreases by 2.04 points. Final average accuracy is 0.54 points lower, final-task accuracy 0.35 points lower, and ECE 0.35 points higher with UniCOP; backward transfer

is approximately unchanged. UniCOP consistently improves future-class OOD ranking across all three orders while maintaining comparable in-domain performance.

To check whether the feature posterior realizes the intended orbit geometry, we examined the posterior after learning the first 50 classes. Each order used the same replay examples for the baseline and UniCOP and 500 feature-posterior samples per learned class. We compared the Gaussian residual, a globally norm-matched Gaussian control, and block-Cayley UniCOP. The norm-matched control changes the global scale but retains the Gaussian sample directions. Table 9 shows that UniCOP permits larger angular displacement while almost eliminating block-energy drift. The corrected geometry aggregation replaces one anomalous baseline run with a serial rerun; the frozen OOD results above come from the separate final campaign.

Table 9: **CLAP4CLIP feature-posterior structure after five sessions.** Means across three held-out class orders. Block-energy drift is relative to the reference feature; lower values indicate better preservation.

Posterior	Median angular displacement	Median block-energy drift
Gaussian residual	35.57°	2.227
Norm-matched Gaussian	35.57°	0.342
UniCOP	72.37°	1.66×10^{-4}

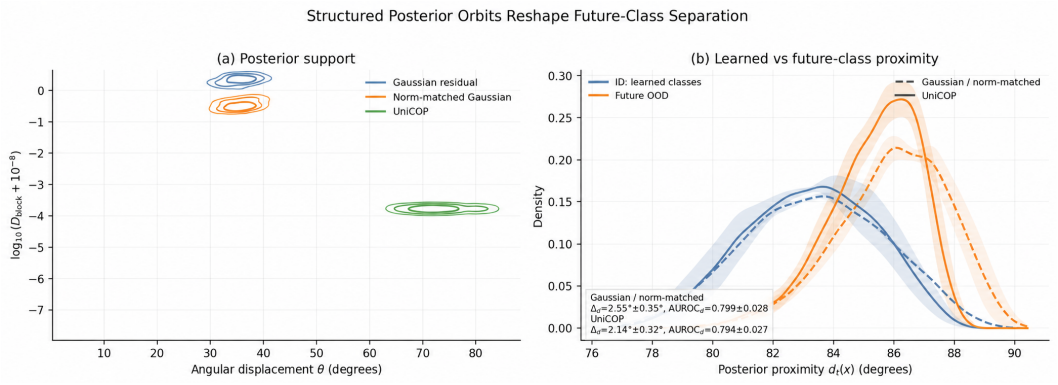


Figure 8: **CLAP4CLIP geometry diagnostic.** The archived plot labels UniCOP as UniCOP. Left: block-Cayley samples preserve feature-block energy while changing direction. Right: a simple mean angular-proximity statistic does not explain the MI OOD gains; its AUROC is 0.794 for UniCOP versus 0.799 for the Gaussian baseline.

A.5 Experiments Across Datasets and Backbones

We evaluate BBB and IVON with and without UniCOP across three dataset–backbone settings: CIFAR-100 with ResNet-20, Tiny-ImageNet with DenseNet-121, and ImageNet-100 with ResNet-50. These experiments assess whether UniCOP remains effective across different data scales and network architectures. Each configuration is trained for 200 epochs with three independent seeds. Table 10 reports the mean and sample standard deviation across seeds.

Table 10 shows that IVON+UniCOP improves all four metrics across the three settings. Its accuracy increases by 0.36 percentage points on CIFAR-100, 0.48 points on Tiny-ImageNet, and 1.12 points on ImageNet-100, accompanied by lower NLL, ECE, and Brier scores. The largest accuracy gain occurs on ImageNet-100 with ResNet-50, where NLL also decreases from 0.714 to 0.693 and ECE from 7.77% to 6.54%.

BBB+UniCOP reduces NLL in all three settings, while its effects on accuracy and calibration depend on the dataset and backbone. On CIFAR-100, accuracy and Brier score improve, but ECE increases. On Tiny-ImageNet, lower NLL accompanies reduced accuracy and higher ECE and Brier score. On ImageNet-100, NLL, ECE, and Brier score improve, while accuracy decreases slightly from 77.29%

Table 10: **In-domain results across datasets and backbones.** Results are reported as mean ± sample standard deviation over three independent seeds after 200 training epochs. Accuracy, ECE, and Brier score are expressed in percent. Higher accuracy and lower NLL, ECE, and Brier score indicate better performance.

Setting	Method	Accuracy(%) ↑	NLL ↓	ECE(%) ↓	Brier(%) ↓
ResNet-20 CIFAR-100	BBB	61.26 ± 1.98	1.565 ± 0.077	9.03 ± 0.40	62.04 ± 1.20
	BBB+**UniCOP**	62.15 ± 0.98	1.525 ± 0.062	15.99 ± 0.74	60.77 ± 1.08
	IVON	64.31 ± 0.42	1.364 ± 0.015	10.29 ± 0.45	48.97 ± 0.38
	IVON+**UniCOP**	64.67 ± 0.40	1.351 ± 0.024	10.21 ± 0.49	48.46 ± 0.56
DenseNet-121 Tiny-ImageNet	BBB	52.41 ± 0.20	4.270 ± 0.012	34.09 ± 0.38	78.14 ± 0.26
	BBB+**UniCOP**	50.45 ± 1.21	4.138 ± 0.062	34.20 ± 0.44	79.85 ± 1.15
	IVON	52.59 ± 0.37	2.123 ± 0.043	14.48 ± 1.19	63.63 ± 0.96
	IVON+**UniCOP**	53.07 ± 0.40	2.107 ± 0.035	14.32 ± 0.41	63.14 ± 0.37
ResNet-50 ImageNet-100	BBB	77.29 ± 1.12	1.011 ± 0.042	10.23 ± 1.13	33.58 ± 1.30
	BBB+**UniCOP**	77.13 ± 0.96	0.913 ± 0.011	8.17 ± 1.09	32.50 ± 0.81
	IVON	80.37 ± 1.04	0.714 ± 0.031	7.77 ± 1.12	30.90 ± 0.62
	IVON+**UniCOP**	81.49 ± 0.90	0.693 ± 0.027	6.54 ± 0.99	29.84 ± 0.75

to 77.13%. These results demonstrate consistent gains for IVON+UniCOP across the tested settings and improvements in predictive likelihood for BBB+UniCOP, with dataset-dependent trade-offs in its other metrics.

A.6 Optimization with Few Training Samples

We examine how the number of posterior samples per training update affects optimization and predictive performance. We sweep $M_{\text{train}} \in \{1, 2, 4, 8, 16, 32\}$ for IVON and IVON+UniCOP on CIFAR-10 with ResNet-20. Each setting uses the same 80/20 training–validation split, batch size of 50, 200 training epochs, and optimizer hyperparameters, with three independent seeds. With a fixed number of updates, the sampled-example forward budget increases linearly from 8 million at $M_{\text{train}} = 1$ to 256 million at $M_{\text{train}} = 32$. We evaluate the learned center predictor to assess the effect of training-time sampling without introducing test-time posterior averaging.

Table 11: **Training-sample sweep on CIFAR-10 with ResNet-20.** All metrics evaluate the learned center predictor. Accuracy, ECE, and Brier score are expressed in percent; higher accuracy and lower NLL, ECE, and Brier score are better.

	IVON				IVON+**UniCOP**			
M_{train}	Acc.(%) ↑	NLL ↓	ECE(%) ↓	Brier(%) ↓	Acc.(%) ↑	NLL ↓	ECE(%) ↓	Brier(%) ↓
1	91.85	0.2934	4.24	12.66	91.90	0.2865	4.07	12.65
2	91.91	0.2956	4.39	12.71	91.95	0.2944	4.19	12.65
4	91.80	0.2907	4.26	12.25	91.92	0.2926	4.40	13.02
8	91.96	0.2982	4.33	12.62	92.12	0.2927	4.24	12.41
16	91.91	0.3058	4.62	12.80	91.93	0.3059	4.47	12.76
32	92.12	0.2965	4.34	12.54	92.19	0.3081	4.28	12.38

Table 11 shows that increasing the training-sample count does not improve all metrics monotonically. IVON achieves its lowest NLL and Brier score with four samples, its lowest ECE with one sample, and its highest accuracy with 32 samples. IVON+UniCOP achieves its lowest NLL and ECE with one sample, while its highest accuracy and lowest Brier score occur with 32 samples. Across all tested budgets, IVON+UniCOP has slightly higher mean accuracy than IVON, although improvements in the other metrics depend on the sample count.

With a single training sample, IVON+UniCOP reaches 91.90% accuracy and improves all four mean metrics over IVON. Increasing its sample count from one to 32 raises accuracy by 0.29 percentage points and reduces Brier score from 12.65% to 12.38%, but increases NLL from 0.2865 to 0.3081 and ECE from 4.07% to 4.28%.