PRECISE BAYESIAN NEURAL NETWORKS VIA SCALE INVARIANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Bayesian neural networks promise calibrated uncertainty, yet the posterior over their weights has been a difficult object: mean-field variational inference underfits it, it predicts well only when cooled below the Bayesian temperature, and sampling it exactly is out of reach at scale. We show that much of this difficulty comes from the parameterisation. Modern networks are normalised, so the length of a weight vector has no effect on the function: the posterior over length is exactly the prior, and every constraint the data impose falls on the direction. We therefore place the posterior on the sphere, as a von Mises–Fisher distribution per unit with a single learned concentration under a uniform prior. In practice the method is adaptive noise: Gaussian noise added at each unit's pre-activation with a scale the network learns, one scalar per unit, whose KL divergence to the uniform prior has a closed-form approximation. Running Hamiltonian Monte Carlo on normalised networks, we find that this family fits: the true directional posterior is close to isotropic within a unit and independent across units, a geometry that layer normalisation and sparse activation create, and the fitted von Mises–Fisher lands on it to within a few degrees, while a factorised Gaussian on the same network spends almost all of its KL on a weight length the data leave undetermined. The learned concentration tracks the HMC posterior's as the dataset grows and follows the width scaling derived from the model; the model works at the untempered evidence lower bound, where weight-space methods typically need cold posteriors; and the learned uncertainty is a signal-to-noise ratio that reads as an angle, comparable across layers of any width.

1 INTRODUCTION

A Bayesian neural network replaces one setting of the weights with a posterior over them and predicts by averaging, with the promise of calibrated uncertainty and a principled account of generalisation (MacKay, 1992; Neal, 1996; Wilson & Izmailov, 2020). The obstacle has been the posterior itself. Mean-field variational inference underfits it and is sensitive to initialisation and hyperparameters (Blundell et al., 2015; Trippe & Turner, 2018; Foong et al., 2020; Farquhar et al., 2020); sampled carefully, it was reported to predict worse than a network trained by SGD unless cooled well below the Bayesian temperature (Wenzel et al., 2020), an effect since traced largely to data augmentation (Izmailov et al., 2021); and establishing what it looks like at all required full-batch Hamiltonian Monte Carlo at a cost that rules the method out for practice. The field has responded with richer variational families (Louizos & Welling, 2016; Sun et al., 2019; Oh et al., 2020; Farquhar et al., 2020), better samplers and tempering, and in practice uncertainty comes from ensembles and dropout (Lakshminarayanan et al., 2017; Ovadia et al., 2019): Bayesian methods reach large models today through post-hoc Laplace approximations (Daxberger et al., 2021; Yang et al., 2024) or a variational learning only recently made to train them (Shen et al., 2024), and the case for Bayesian deep learning at scale is still being argued (Papamarkou et al., 2024).

We start from a property of the networks themselves. Modern architectures are normalised: batch, layer and weight normalisation each remove the scale of a weight vector from the computation, so the function is unchanged when a weight vector is multiplied by any positive constant (Ioffe & Szegedy, 2015; Ba et al., 2016; Salimans & Kingma, 2016; Meng et al., 2019). The likelihood is therefore flat along the length of every weight vector. Under an isotropic prior the posterior factorises exactly: its marginal over length is the prior, and everything the data determine is a direction on the

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.

unit sphere. A Gaussian posterior in weight space ties the two together: the same variances set its radial shell and its angular spread, so fitting the direction to the data distorts the length. Farquhar et al. (2020) identified the sampling pathology this produces; Section 4 shows what it does to the fit.

We therefore place the weights on the sphere. Each weight vector is a unit vector, its prior is uniform on the sphere and its approximate posterior is a von Mises–Fisher distribution with a mean direction and one concentration κ. The KL divergence to the uniform prior has a closed-form approximation in the standard deviation σ of additive Gaussian noise at the unit's pre-activation, with no Bessel functions, so the layer is adaptive noise: Gaussian noise at each pre-activation with a scale the network learns, and a KL term that sets it. We call it a concentration-adapted perturbation (CAP). It costs one scalar per unit, or per layer, where mean-field and variational dropout carry one per weight; a network that already injects noise becomes Bayesian by making that noise adaptive under the KL term. The learned σ converts through κ to the mean angle between a sampled direction and the mean direction; that angle does not depend on fan-in, so the uncertainty of a 27-input layer and a 4096-input layer are read on one scale, on which 90° is the uniform prior.

The family fits the true posterior: where HMC converges the fitted von Mises–Fisher sits on the true angular law, and on a real network its concentration follows the true one as data grow (Section 4). The KL is one closed-form expression per unit that tracks the exact divergence to within a near-constant factor over the operating range (Section 2). The model works at its Bayesian objective, $\beta = 1$ (Section 6). The uncertainty adapts across architecture and dataset size and stays calibrated as it does (Section 5). And it can be read, as a signal-to-noise ratio and an angle on an absolute scale (Section 2).

2 Directional posteriors for normalised networks

Scale invariance. The starting point is an invariance that every normalised network has and that variational inference over weights has rarely used. Consider a unit that computes $u = w^\top x + b$ and is followed by a normalisation that removes scale: batch normalisation without an affine transform, layer normalisation, or weight normalisation with a fixed gain. The network's output is then the same for w and cw for every $c > 0$, and so is the likelihood, $p(\mathcal{D} \mid cw) = p(\mathcal{D} \mid w)$.

Proposition 1 (factorisation)**.** *Let the likelihood be invariant to $w \mapsto cw$ for $c > 0$, and let the prior factorise into a law over length and a law over direction, $p(w) = p_r(\|w\|)\, p_\theta(w/\|w\|)$. Then the posterior factorises the same way, $p(w \mid \mathcal{D}) = p_r(\|w\|)\, p_\theta(\theta \mid \mathcal{D})$ with $p_\theta(\theta \mid \mathcal{D}) \propto p_\theta(\theta)\, p(\mathcal{D} \mid \theta)$.*

The proof is one line: the likelihood is a function of the direction alone, so it multiplies the directional factor and leaves the radial one untouched. The data can rotate a weight vector but cannot lengthen or shorten it, and any variance a posterior places along the length is unconstrained by the data. Every isotropic prior is of this form, and for the isotropic Gaussian $\mathcal{N}(0, s^2 I)$ the directional factor is the uniform distribution on S^{D-1}: the uniform prior on the sphere is the part of the standard Gaussian prior that the likelihood can see (Fortuin et al., 2022). A factorised Gaussian $q(w) = \mathcal{N}(\mu, \sigma^2 I)$ couples the two: for a sample $w = \mu + \sigma\varepsilon$ both the squared length and the cosine to μ are set, in high dimension, by the single ratio $\sigma\sqrt{D}/\|\mu\|$, so tightening the direction also tightens the length. Farquhar et al. (2020) identified the sampling consequences of the resulting shell; Section 4 shows that at large fan-in almost all of the Gaussian's KL is spent on the radius. Fig. 1 draws the three priors in two dimensions.

The layer. Each unit's weight is a unit vector $w \in S^{D-1}$, its prior is uniform, and its approximate posterior is a von Mises–Fisher distribution

$$q(w \mid \mu, \kappa) = C_D(\kappa) \exp(\kappa\, \mu^\top w), \qquad \|\mu\| = 1, \quad \kappa \geq 0, \tag{1}$$

with mean direction μ, the network's weight, and concentration κ, one scalar per unit or per layer (Mardia & Jupp, 2000; Banerjee et al., 2005). Under q the pre-activation has mean $A_D(\kappa)\, \mu^\top x$, where $A_D(\kappa) = I_{D/2}(\kappa)/I_{D/2-1}(\kappa)$ is the mean resultant length, and variance approximated by $\|x\|^2/(\kappa + D)$, which is exact in both limits (Appendix A). Writing $\sigma_{\text{eff}} = \sigma_u / A_D(\kappa)$ for the noise relative to the signal, the layer computes

$$u = \hat{\mu}^\top x + b + \sigma_{\text{eff}}\, \text{rms}(x)\, \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, 1), \tag{2}$$

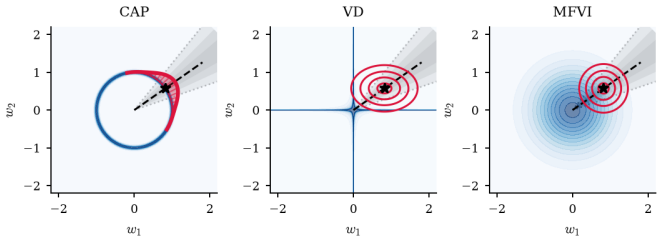

Figure 1: Prior geometry for CAP, VD and MFVI at fan-in 2: the likelihood is the grey wedge, constant along the radial direction, and the star is the posterior mode. CAP's prior is uniform on the circle and its posterior a cap on it; VD's per-weight log-uniform prior, $p(w) \propto \prod_i 1/|w_i|$, diverges on the coordinate axes, origin included, where the likelihood is finite; MFVI's isotropic Gaussian, whose directional marginal is CAP's prior, spends mass along the radius.

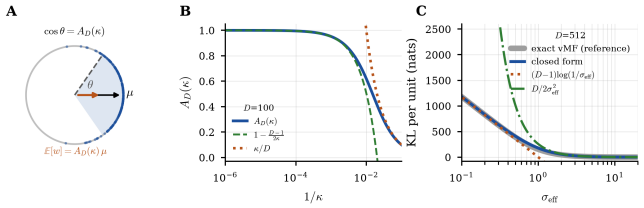

Figure 2: The family and its KL. (A) A von Mises–Fisher on the circle: samples on a cap around μ, whose mean $A_D(\kappa)\mu$ is shorter than μ by the factor that sets the typical angle, $\cos\theta = A_D(\kappa)$. (B) $A_D(\kappa)$ against $1/\kappa$ at $D = 100$ with its two asymptotes. (C) KL per unit against σ_{eff} at $D = 512$: the closed form of Eq. (3) (blue) over the exact von Mises–Fisher KL (grey), with both limits of the closed form drawn over the full range; the y axis is linear so that where each asymptote departs is visible.

and learns $\log \sigma_{\text{eff}}$. The factor A_D multiplies signal and noise together, so a downstream normalisation absorbs it (Appendix A). The implementation is therefore an additive-noise layer, a close relative of Gaussian dropout (Srivastava et al., 2014; Wang & Manning, 2013); the sphere enters through the KL and the reading of σ_{eff}. Fig. 2A draws the family.

The dictionary between noise and angle. Because $\sigma_{\text{eff}} = \sqrt{D/(\kappa + D)}\,/A_D(\kappa)$ and $\mathbb{E}_q \cos\angle(w, \mu) = A_D(\kappa)$, the learned scale is a signal-to-noise ratio that converts monotonically to the angle between a draw and the mean direction. For large D both this angle and the KL per unit depend on κ and D only through κ/D (Table 4, Fig. 8, Appendix B), so a σ_{eff} read in layers of any fan-in is compared on one scale on which $\sigma_{\text{eff}} = 0.33$ is 19°, $\sigma_{\text{eff}} = 10$ is 85° and 90° is the uniform prior. The α of variational dropout and the σ of a Gaussian posterior both scale with the weight magnitude and have no such reading.

The KL divergence. The KL from the von Mises–Fisher to the uniform prior is $\kappa A_D(\kappa) + \log C_D(\kappa) + \log |S^{D-1}|$ (Sra, 2016): exact, but through Bessel functions of order $D/2$, which is the term Oh et al. (2020) have to bound numerically. In the σ_{eff} parameterisation it is approximated, for a layer of M units with fan-in D, by

$$\text{KL} = \frac{M(D-1)}{2} \log\Big(1 + \frac{D}{D-1}\,\sigma_{\text{eff}}^{-2}\Big), \tag{3}$$

which has the correct behaviour at both ends (Fig. 2C). Its relative error against the exact divergence depends on κ/D alone: negligible once κ/D is in the tens, and about 20% near $\kappa/D = 1$, where our Fashion-MNIST runs sit, which in training acts as a fixed rescaling of the KL weight (Appendix B). The objective is $\mathcal{L} = \text{NLL} + \beta \sum_l \text{KL}_l$; at test time one averages S stochastic passes or uses the mean direction alone (Section 6). Other priors on the sphere add one term to the KL; an ℓ_1 prior gives fan-in sparsity from the direction alone (Appendix F).

The family makes two further assumptions, that the posterior is isotropic within a unit and that it factorises across units. Section 3 tests both against HMC samples.

Table 1: The two factorisation assumptions under HMC ($D = 40$, $K = 6$): mean $|\rho|$ between units, mean $|r|$ between tangent coordinates within a unit and the AM/HM overconfidence of Proposition 2, with and without layer normalisation and sparse activation (a negative bias in the toy).

condition	cross-unit $\lvert\rho\rvert$	within-unit $\lvert r\rvert$	AM/HM	$\hat{R}$	ESS
erf, $b = 0$	0.102	0.197	1.888	1.056	245
erf, $b = 0$, + LN	0.087	0.033	1.043	1.000	3104
ReLU, $b = -2$	0.029	0.090	1.424	1.079	114
ReLU, $b = -2$, + LN	0.031	0.039	1.077	1.000	6000

3 THE POSTERIOR UNDER NORMALISATION

Within a unit the tangent covariance of a von Mises–Fisher is $(1/\kappa)\, I$, isotropic; across units q factorises. Most scalable Bayesian neural networks make the same two assumptions in weight space, where the posterior is too hard to sample to test them. On the directions of a small normalised network HMC converges, so both can be measured, and each turns out to depend on a property of the trained network: whether it is normalised, and how sparsely its units fire.

Anisotropy within a unit. Near its mode the true directional posterior is Gaussian on the tangent space with some precision H of dimension $n = D-1$. A von Mises–Fisher must summarise H with a single number, and the ELBO's stationary point in κ picks the arithmetic mean of H's eigenvalues, while reproducing the true angular spread would need the harmonic mean.

Proposition 2 (anisotropy)**.** *At the mode, $\kappa_q = \mathrm{tr}(H)/n$ and $\kappa_{\mathrm{true}} = n/\mathrm{tr}(H^{-1})$, so $\kappa_q/\kappa_{\mathrm{true}} = \mathrm{tr}(H)\,\mathrm{tr}(H^{-1})/n^2 \geq 1$, with equality if and only if H is isotropic.*

The inequality is Cauchy–Schwarz (Appendix C). A von Mises–Fisher is therefore overconfident by exactly the ratio of the arithmetic to the harmonic mean of the tangent Hessian's spectrum, and calibrated when the spectrum is flat; the prediction gives 1.0 for an erf teacher and 2.3 for a ReLU teacher with a negative bias, and HMC measures 0.98 and 2.3. The error of a single concentration per unit is thus a known function of the posterior's shape, independent of dimension over the eightfold range of D we checked, and it vanishes when the Hessian is flat, which is the condition normalisation produces below.

Measurements. We sample the posterior of a student network with $D = 40$ inputs and $K = 6$ units by HMC and report the mean absolute correlations between units and between tangent coordinates within a unit, the AM/HM ratio of Proposition 2 and the sampler's diagnostics (Table 1). Two factors are crossed: layer normalisation, and the sparsity of the units, which the toy controls through the teacher's bias (a negative bias makes each unit fire on a small region disjoint from the others); in a trained network the bias is learned, and this sparsity arises on its own in ReLU networks or can be induced with an activity penalty.

The two factors affect different assumptions. Sparse activation reduces the cross-unit correlation about threefold but leaves the within-unit correlation above its normalised level (0.090 against 0.033). Layer normalisation reduces the within-unit correlation sixfold, brings the AM/HM ratio to within a few percent of one and lets the sampler mix, raising the effective sample size by an order of magnitude; it leaves the cross-unit correlation where it was (Fig. 3; all four conditions in Fig. 7). Sparsity matters to a posterior because units that fire on the same inputs can substitute for one another, so the likelihood constrains only their combination and their directions are correlated; once their firing regions are disjoint each direction is constrained on its own. The fraction of one unit's rotation that the others can compensate, an R^2 we measure directly, falls from nearly one to nearly zero as the units become sparse, and the overconfidence of the factorised fit falls fivefold with it. The two effects are distinct: with a single unit there is no other unit to correlate with, and the overconfidence is still 2.3, the anisotropy value. The correlation can also be modelled, with an equicorrelated q whose KL gains $-\frac{D-1}{2}\log\det R$; learning it takes the erf teacher's overconfidence from eightfold to nearly calibrated, at the cost of one learned correlation and a $K \times K$ Cholesky per step (Appendix E).

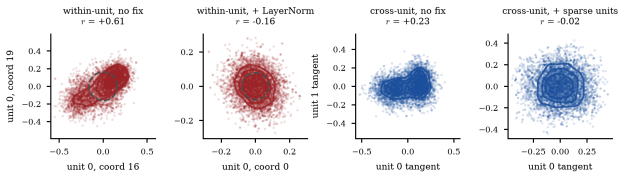

Figure 3: Correlations in the true posterior with and without each factor, HMC samples for the erf teacher. Left pair: two weight coordinates of one unit in the coordinate basis, with the dashed circle a single concentration assumes; $r = +0.61$ without normalisation, -0.16 with layer normalisation (one strongly correlated pair; Table 1 gives the mean over pairs). Right pair: a tangent coordinate of unit 0 against one of unit 1; $r = +0.23$ with dense units, -0.02 with sparse units.

Normalisation therefore has two separate effects: it removes the length coordinate, which makes the spherical family appropriate, and it flattens the tangent Hessian and improves mixing, which makes a single concentration per unit adequate; sparse activation, which normalised ReLU networks develop as they train, takes care of the factorisation across units.

4 Comparison with the true posterior

An approximation is judged by its distance from what it approximates, which in weight space has rarely been measured because the reference is out of reach (Izmailov et al., 2021; Sommer et al., 2024). On the sphere it is within reach, on a small problem where HMC can be run to convergence and on Fashion-MNIST.

The toy. A comparison against HMC is informative only where the posterior is determined by the data and the chains have mixed, so we sweep teacher bias against dataset size on a small student, run HMC (Neal, 2011; Hoffman & Gelman, 2014) at each setting, and keep the settings whose chains converge by split-$\hat{R}$ (Vehtari et al., 2021) and whose posterior is determined well away from uniform (Appendix G); since near the uniform prior any family fits. At the two settings that pass the true posterior is determined to about $14°$, and the von Mises–Fisher fit sits on the true angular law to within $2.2°$ and $3.1°$ in 1-Wasserstein distance (Fig. 6, Appendix J). Its remaining error is the one Proposition 2 predicts: $1.5\times$ overconfident in κ, matching the $1.4\times$ we measure on Fashion-MNIST below.

The radial coordinate. Under a batch-normalised pre-activation the true radial marginal is the prior in closed form, so the KL of a factorised Gaussian to the prior splits exactly into a part along $\hat{\mu}$, the radial direction the likelihood cannot see, and a part along the $D - 1$ tangent directions it can. Across a $32\times$ range of fan-in with labels per parameter held fixed, the radial share of MFVI's KL rises from about half to nearly all of it, while CAP's is zero because the family has no radial coordinate; over the same range CAP's angular error to the teacher stays near $9°$ while MFVI's grows from $12°$ to $49°$ (Fig. 5, Appendix J). Both grow with D. Fig. 12 compares the KL each family pays for the noise it actually injects.

Fashion-MNIST. We sample the directional posterior of a network with one stochastic layer of 32 units and fan-in 512, under CAP's own uniform prior and with multi-chain diagnostics (Appendix I). Table 2 gives the learned and the true concentration for each family at the full dataset, and the same protocol at three smaller dataset sizes gives the dependence on data.

CAP's concentration is $1.4\times$ the true one, as Proposition 2 predicts and as in the toy; variational dropout's is about ten percent below it; the mean-field Gaussian's is less than half of it. Across dataset sizes the learned concentrations of CAP and of variational dropout both track the true one within tens of percent, and mean-field VI is off by a constant factor of three to four. And over the three dataset sizes the true posterior concentrates roughly as $N^{1/4}$, against N^1 for an identifiable direction: the posterior of a real network concentrates far more slowly than the derivation of Section 5 assumes, and CAP follows the measured rate, with a slightly steeper slope, which is why it drifts from mildly under- to mildly over-confident across the range.

Table 2: Learned against true concentration on Fashion-MNIST at $N = 60{,}000$ (left) and the slopes of $\log \kappa$ against $\log N$ over $N = 2000$ to $50{,}000$ with the ratio at the ends (right); κ_{HMC} is measured in each method's own basin.

	κ_q	κ_{HMC}	ratio	$\hat{R}$	ESS	slope κ_q	slope κ_{HMC}	ratio at 2k	at 50k
CAP	388	273	1.42	1.007	854	$+0.345$	$+0.217$	0.86	1.30
VD	397	448	0.89	1.002	3756	$+0.220$	$+0.287$	0.99	0.80
MFVI	130	332	0.39	1.004	1429	$+0.339$	$+0.261$	0.26	0.34

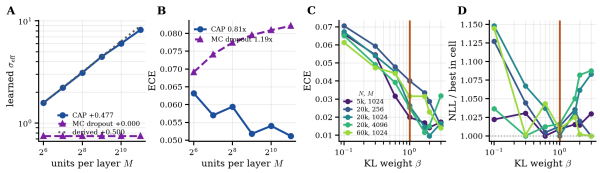

Figure 4: Fashion-MNIST, first block deterministic. (A) Learned σ_{eff} against units per layer at fixed fan-in 256, with the derived $M^{1/2}$ line; CAP measures $+0.477$, MC dropout has no parameter to learn. (B) ECE over the same widths from two arms that start level at $M = 64$: CAP's falls by a fifth, MC dropout's rises by a fifth (all arms at the same realised initial noise, one noise-free warm-up epoch, MC dropout at its validation-tuned rate). (C) ECE against the KL weight for five (N, M) cells; the optimum is at or above $\beta = 1$ in every cell. (D) Test NLL relative to the best in each cell; $\beta = 1$ costs at most 1.6%.

5 Predicted Scaling

A posterior should respond to changes in the problem in predictable ways, and the model predicts three: the concentration of a unit should grow with the data that constrain it and be indifferent to its fan-in, the noise a layer carries should grow with its number of units, and the overconfidence of Proposition 2 should vanish as the layer becomes identifiable. None of these predictions has a free parameter.

Width. The noise injected into the M units of a layer is independent across units while the signal they carry is coherent, so the signal-to-noise ratio of the layer is $\sqrt{M}$ times that of a unit, and holding the layer's effective regularisation fixed as M grows requires the per-unit noise to grow as $\sigma^*_{\mathrm{eff}} \propto M^{1/2}$ (Appendix D). We test this on Fashion-MNIST with the measured layer's fan-in pinned and its unit count the only thing that varies over a $32\times$ range, with nothing retuned at any width. A posterior quantity should follow the square root without retuning; the fitted exponent is 0.48 against the derived 0.50 (Fig. 4A). The noise level therefore transfers across widths, and calibration improves as the layer widens, where MC dropout, whose rate is fixed and which starts level with CAP at the narrowest width, loses calibration as the layer widens (Fig. 4B).

Fan-in and data. At the stationary point the concentration of a unit is $\kappa^* \approx NC$, with C the likelihood's curvature along the unit's direction and no dependence on D; over a $32\times$ range of fan-in the exponent of κ in D is 0.1 for CAP against a predicted 0, and 1.0 for variational dropout against its predicted 1. Where a direction is identifiable $\kappa \propto N$ and the angle falls as $N^{-1/2}$; on a teacher calibrated to a fixed Bayes accuracy it falls with slope -0.52 and -0.54 at two step budgets against the predicted -0.5, with no drift in overconfidence, and a correlated teacher flattens it to -0.39, consistent with Section 3. Where the premise fails, on Fashion-MNIST, the true concentration grows roughly as $N^{1/4}$ because the likelihood's curvature collapses as the training set is fit (Appendix J), so the $N^{-1/2}$ law is an upper bound on how fast a real posterior concentrates, and CAP tracks the actual rate; its calibration across the same range of N is in Fig. 13.

Overconfidence. The AM/HM ratio of Proposition 2 vanishes with information, falling as the inverse of the information per weight as the Hessian's spectrum flattens, so it is a finite-sample

effect; it also falls with width, from 2.5 to 1.7 as a layer grows from four to thirty-two units, until past $K/D \approx 0.3$ the problem stops being identifiable and more data make it worse (Appendix J).

6 EXPERIMENTS

Table 3: Fashion-MNIST at width 1024, all six arms under one protocol: noise on the same two blocks, every learned-noise arm initialised at the same realised noise-to-signal ratio 0.1, one noise-free warm-up epoch, $S = 16$ samples at evaluation, one seed. MC dropout is shown at its validation-tuned rate $p = 0.3582$; with no noise parameter it would otherwise keep the matched initial ratio. $\beta = 0.3$ for every variational arm.

	accuracy	ECE	NLL	train−test gap
deterministic	0.8631	0.1033	0.7679	0.1310
MFVI	0.8518	0.0846	0.5960	0.1218
MC dropout	0.8700	0.0809	0.5660	0.1232
VD	0.8640	0.0686	0.5115	0.1209
CAP	0.8605	0.0540	0.4731	0.1276
Radial	0.8653	0.0296	0.4525	0.1105

Fashion-MNIST at width 1024. Every arm that carries noise beats the deterministic network on both calibration and likelihood (Table 3). CAP is second on both, behind Farquhar's radial posterior; variational dropout, MC dropout and the mean-field Gaussian trail, in that order on both. The accuracies are within two points of each other (0.852 to 0.870), so the ordering reflects uncertainty. Predictive accuracy converges well before $S = 64$ samples for every arm. On the UCI regression benchmark, under the same code, CAP takes the best RMSE on all four tuned datasets and the best test log-likelihood on two (Appendix H).

The KL weight. The untempered evidence lower bound has $\beta = 1$. Fig. 4C,D sweeps β on Fashion-MNIST across a $12\times$ range of dataset size and a $16\times$ range of width: in every cell the calibration optimum is at or above $\beta = 1$, and $\beta = 1$ is within 2% of the best test likelihood (a finer sweep of one cell is in Fig. 11). Weight-space posteriors are typically tempered by one to three orders of magnitude (Wenzel et al., 2020); on the sphere no tempering was needed in any cell we ran. The result depends on the input block being deterministic, as it is throughout (a partially stochastic network in the sense of Sharma et al., 2023): putting CAP on the raw-pixel input layer as well drags the optimum down to $\beta = 0.1$, while whitening the inputs leaves it at 1, so input correlation is not the cause. Against a known teacher, where calibration can be measured directly, the fit is calibrated near $\beta = 1$ and about $27\times$ overconfident at $\beta = 0.1$, while the realised angle to the teacher stays at about 13°, so β changes only the reported uncertainty (Fig. 10). The model therefore works at its Bayesian objective, $\beta = 1$; on the small-data UCI regression benchmark the selected β is an order of magnitude lower (Appendix H).

High noise and deterministic evaluation. When fan-in is allowed to grow with width, σ_{eff} absorbs a $26\times$ increase in noise, taking the directional uncertainty from about 40° to nearly 90° with accuracy flat and calibration improved: a unit determined to within a few degrees of uniform works because the prediction averages over the posterior and the layer's signal-to-noise ratio is $\sqrt{M}$ times the unit's. Noise levels this high have also been reported for validation-tuned noise injection (Brito, 2025). A CAP network can also be evaluated with the noise off: the mean direction loses a fifth of the accuracy under the batch-norm statistics collected with the noise on, and re-estimating them with the noise off, two forward passes and no change to any weight, restores it, whereas the mean of a factorised Gaussian posterior stays 28 points below its own Monte Carlo accuracy (Appendix I). At deployment a CAP network therefore costs what a deterministic network costs, with S passes when calibrated probabilities are needed.

7 Related work

Weight-space inference. Hinton & van Camp (1993), Graves (2011) and Blundell et al. (2015) put a factorised Gaussian over the weights; later variants reduce its variance (Wu et al., 2019; Osawa et al., 2019) and train large networks (Shen et al., 2024). Its documented problems are over-pruning (Trippe & Turner, 2018), limited expressiveness (Foong et al., 2020) and the soap bubble (Farquhar et al., 2020); the cold-posterior effect and its attribution to prior, curation and augmentation (Wenzel et al., 2020; Aitchison, 2021; Izmailov et al., 2021; Nabarro et al., 2022; Kapoor et al., 2022) are findings about the weight-space posterior itself. The response has been richer families (Louizos & Welling, 2016; Sun et al., 2017; Louizos & Welling, 2017; Sun et al., 2019; Rudner et al., 2022) or post-hoc Laplace approximations (Ritter et al., 2018; Daxberger et al., 2021; Kristiadi et al., 2020); we take the opposite approach, removing a coordinate from the model so that a one-parameter family fits. Gelberg et al. (2024) show that permutation symmetries bias variational posteriors unless the family respects them, and the scale symmetry is the analogous case; Sharma et al. (2023) show a network need not be stochastic in every layer, and our deterministic input block is one instance.

Spheres and noise. Farquhar et al. (2020) place the perturbation in hyperspherical coordinates with a uniform, unlearned angular distribution and per-weight parameters; CAP puts the sphere on the weight, learns the concentration and has fewer variational parameters than mean-field by a factor of the fan-in. Oh et al. (2020) are the closest neighbour, a von Mises–Fisher direction with one concentration per layer and a magnitude with a horseshoe-type prior (Louizos et al., 2017), through Bessel ratios bounded numerically; under normalisation the magnitude is unidentifiable, so CAP drops it and the KL loses its Bessel functions. Ghoshal & Tucker (2021) constrain Bayes-by-backprop weights to the sphere without a closed-form KL; von Mises–Fisher and Bingham distributions appear elsewhere as latents and outputs (Davidson et al., 2018; Kumar & Tsvetkov, 2019; Gilitschenski et al., 2020); variational dropout (Kingma et al., 2015; Molchanov et al., 2017) obtains sparsity from a log-uniform prior on the magnitude that makes the posterior improper (Hron et al., 2018). Normalisation (Salimans & Kingma, 2016; Ioffe & Szegedy, 2015; Ba et al., 2016; Meng et al., 2019) was proposed for optimisation, and under it and weight decay only the angular update matters (Hoffer et al., 2018; Wan et al., 2021; Kodryan et al., 2022; Kosson et al., 2024), while recent architectures put weights on the sphere outright (Karras et al., 2024; Loshchilov et al., 2025) and recent optimisers orthogonalise the update of each weight matrix (Jordan et al., 2024; Liu et al., 2025); Section 3 measures what the same invariance does to the posterior. Gaussian and MC dropout (Srivastava et al., 2014; Wang & Manning, 2013; Gal & Ghahramani, 2016; Gal et al., 2017) and cross-regularisation (Brito, 2025) inject noise where CAP does; CAP gives that noise a posterior and a KL.

8 Discussion

Learning in a normalised network is angle fitting: gradient descent on the mean direction moves a point on a sphere, and the posterior is the angular uncertainty of that fit, on a scale where a layer's uncertainty means the same thing at any fan-in, 90° is the prior, and the $\sigma_{\text{eff}} = 8.2$ our widest layer reaches is a direction determined to 83°. All of it follows from removing the length coordinate: the family fits the true posterior where HMC can be run and follows its concentration on a real network; the KL is one closed-form expression per unit that tracks the exact divergence to within a near-constant factor; the model works at its Bayesian objective, with none of the KL down-weighting weight-space inference relies on, because the KL on the sphere has no radial term, which is where most of MFVI's KL goes (Section 4); the uncertainty adapts across width and dataset size and stays calibrated; and it can be read as a signal-to-noise ratio and an angle.

The closed form overestimates the exact divergence by about a fifth near $\kappa/D = 1$, and because that overestimate is nearly constant across the range a trained network visits (Fig. 9), it acts as a rescaling of β. The small-data UCI benchmark selects a β an order of magnitude below the classification value.

The evidence here comes from multilayer perceptrons and one small convolutional network (Fig. 14), on Fashion-MNIST, CIFAR-10 and the UCI regressions. The invariance the method rests on holds in any architecture that normalises away the scale of a weight vector, which includes normalised transformers such as nGPT (Loshchilov et al., 2025). Transformers and non-image modal-

ities are outside the experiments here. Constraining the weights to the sphere leaves a normalised network's function class unchanged, since the function depends on direction alone. With the length removed the posterior is simpler than in weight space, and a von Mises–Fisher with one concentration per unit fits it to within a few degrees, in any network that is normalised.

AI USE STATEMENT

We used a large language model as an assistant throughout this work: for drafting and revising the text, for parts of the experimental code and for working through the derivations. The authors verified and take responsibility for the final content.

REPRODUCIBILITY STATEMENT

Appendix I gives the training protocol and Appendix H the UCI protocol. Derivations for every proposition are in Appendices A to F. Code will be released with the camera-ready version.

REFERENCES

Laurence Aitchison. A statistical theory of cold posteriors in deep neural networks. In *International Conference on Learning Representations*, 2021.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Arindam Banerjee, Inderjit S. Dhillon, Joydeep Ghosh, and Suvrit Sra. Clustering on the unit hypersphere using von Mises–Fisher distributions. *Journal of Machine Learning Research*, 6:1345–1382, 2005.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. In *International Conference on Machine Learning*, 2015.

Carlos Stein Brito. Cross-regularization: Adaptive model complexity through validation gradients. In *International Conference on Machine Learning*, 2025.

Tim R. Davidson, Luca Falorsi, Nicola De Cao, Thomas Kipf, and Jakub M. Tomczak. Hyperspherical variational auto-encoders. In *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 856–865, 2018.

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*, 2021.

Sebastian Farquhar, Michael A. Osborne, and Yarin Gal. Radial Bayesian neural networks: Beyond discrete support in large-scale Bayesian deep learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 1352–1362. PMLR, 2020.

Andrew Y. K. Foong, David R. Burt, Yingzhen Li, and Richard E. Turner. On the expressiveness of approximate inference in Bayesian neural networks. In *Advances in Neural Information Processing Systems*, 2020.

Vincent Fortuin, Adrià Garriga-Alonso, Sebastian W. Ober, Florian Wenzel, Gunnar Rätsch, Richard E. Turner, Mark van der Wilk, and Laurence Aitchison. Bayesian neural network priors revisited. In *International Conference on Learning Representations*, 2022.

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.

Yarin Gal, Jiri Hron, and Alex Kendall. Concrete dropout. *Advances in Neural Information Processing Systems*, 30, 2017.

Yoav Gelberg, Tycho F. A. van der Ouderaa, Mark van der Wilk, and Yarin Gal. Variational inference failures under model symmetries: Permutation invariant posteriors for Bayesian neural networks. In *Proceedings of the Geometry-grounded Representation Learning and Generative Modeling Workshop (GRaM)*, volume 251 of *Proceedings of Machine Learning Research*, pp. 233–248, 2024.

Biraja Ghoshal and Allan Tucker. Hyperspherical weight uncertainty in neural networks. In *Advances in Intelligent Data Analysis XIX*, pp. 3–11. Springer, 2021.

Igor Gilitschenski, Roshni Sahoo, Wilko Schwarting, Alexander Amini, Sertac Karaman, and Daniela Rus. Deep orientation uncertainty learning based on a Bingham loss. In *International Conference on Learning Representations*, 2020.

Alex Graves. Practical variational inference for neural networks. In *Advances in Neural Information Processing Systems 24*, pp. 2348–2356. Curran Associates, Inc., 2011.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. *International Conference on Machine Learning*, pp. 1321–1330, 2017.

José Miguel Hernández-Lobato and Ryan P. Adams. Probabilistic backpropagation for scalable learning of Bayesian neural networks. In *International Conference on Machine Learning*, 2015.

Geoffrey E. Hinton and Drew van Camp. Keeping the neural networks simple by minimizing the description length of the weights. In *Conference on Computational Learning Theory*, 1993.

Elad Hoffer, Ron Banner, Itay Golan, and Daniel Soudry. Norm matters: Efficient and accurate normalization schemes in deep networks. In *Advances in Neural Information Processing Systems*, 2018.

Matthew D. Hoffman and Andrew Gelman. The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15, 2014.

Jiri Hron, Alexander G. de G. Matthews, and Zoubin Ghahramani. Variational Bayesian dropout: pitfalls and fixes. In *International Conference on Machine Learning*, 2018.

S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the International Conference on Machine Learning*, 2015.

Pavel Izmailov, Sharad Vikram, Matthew D. Hoffman, and Andrew Gordon Wilson. What are Bayesian neural network posteriors really like? In *International Conference on Machine Learning*, pp. 4629–4640, 2021.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. Blog post.

Sanyam Kapoor, Wesley J. Maddox, Pavel Izmailov, and Andrew Gordon Wilson. On uncertainty, tempering, and data augmentation in Bayesian classification. In *Advances in Neural Information Processing Systems*, 2022.

Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and improving the training dynamics of diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.

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

Maxim Kodryan, Ekaterina Lobacheva, Maksim Nakhodnov, and Dmitry Vetrov. Training scale-invariant neural networks on the sphere can happen in three regimes. In *Advances in Neural Information Processing Systems*, 2022.

Atli Kosson, Bettina Messmer, and Martin Jaggi. Rotational equilibrium: How weight decay balances learning across neural networks. In *International Conference on Machine Learning*, 2024.

Agustinus Kristiadi, Matthias Hein, and Philipp Hennig. Being Bayesian, even just a bit, fixes overconfidence in ReLU networks. In *International Conference on Machine Learning*, 2020.

Sachin Kumar and Yulia Tsvetkov. Von Mises–Fisher loss for training sequence to sequence models with continuous outputs. In *International Conference on Learning Representations*, 2019.

Gerhard Kurz and Uwe D. Hanebeck. Stochastic sampling of the hyperspherical von Mises–Fisher distribution without rejection methods. In *IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, pp. 1–6, 2015.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 30, 2017.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, Yanru Chen, Huabin Zheng, Yibo Liu, Shaowei Liu, Bohong Yin, Weiran He, Han Zhu, Yuzhi Wang, Jianzhou Wang, Mengnan Dong, Zheng Zhang, Yongsheng Kang, Hao Zhang, Xinran Xu, Yutao Zhang, Yuxin Wu, Xinyu Zhou, and Zhilin Yang. Muon is scalable for LLM training. *arXiv preprint arXiv:2502.16982*, 2025.

Ilya Loshchilov, Cheng-Ping Hsieh, Simeng Sun, and Boris Ginsburg. nGPT: Normalized transformer with representation learning on the hypersphere. In *International Conference on Learning Representations*, 2025.

Christos Louizos and Max Welling. Structured and efficient variational deep learning with matrix Gaussian posteriors. In *International Conference on Machine Learning*, 2016.

Christos Louizos and Max Welling. Multiplicative normalizing flows for variational Bayesian neural networks. In *International Conference on Machine Learning*, pp. 2218–2227, 2017.

Christos Louizos, Karen Ullrich, and Max Welling. Bayesian compression for deep learning. In *Advances in Neural Information Processing Systems*, 2017.

David J. C. MacKay. A practical Bayesian framework for backpropagation networks. *Neural Computation*, 4(3):448–472, 1992.

Kanti V. Mardia and Peter E. Jupp. *Directional Statistics*. John Wiley & Sons, 2000.

Qi Meng, Shuxin Zheng, Huishuai Zhang, Wei Chen, Zhi-Ming Ma, and Tie-Yan Liu. G-SGD: Optimizing ReLU neural networks in its positively scale-invariant space. In *International Conference on Learning Representations (ICLR)*, 2019.

Dmitry Molchanov, Arsenii Ashukha, and Dmitry Vetrov. Variational dropout sparsifies deep neural networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2498–2507. PMLR, 2017.

Seth Nabarro, Stoil Ganev, Adrià Garriga-Alonso, Vincent Fortuin, Mark van der Wilk, and Laurence Aitchison. Data augmentation in Bayesian neural networks and the cold posterior effect. In *Uncertainty in Artificial Intelligence*, 2022.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using Bayesian binning. In *AAAI Conference on Artificial Intelligence*, 2015.

Radford M. Neal. *Bayesian Learning for Neural Networks*. Springer, 1996.

Radford M. Neal. MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC, 2011.

Changyong Oh, Kamil Adamczewski, and Mijung Park. Radial and directional posteriors for bayesian deep learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04): 5298–5305, 2020.

Kazuki Osawa, Siddharth Swaroop, Anirudh Jain, Runa Eschenhagen, Richard E. Turner, Rio Yokota, and Mohammad Emtiyaz Khan. Practical deep learning with Bayesian principles. In *Advances in Neural Information Processing Systems*, 2019.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Theodore Papamarkou, Maria Skoularidou, Konstantina Palla, Laurence Aitchison, Julyan Arbel, David Dunson, Maurizio Filippone, Vincent Fortuin, Philipp Hennig, José Miguel Hernández-Lobato, Aliaksandr Hubin, Alexander Immer, Theofanis Karaletsos, Mohammad Emtiyaz Khan, Agustinus Kristiadi, Yingzhen Li, Stephan Mandt, Christopher Nemeth, Michael A. Osborne, Tim G. J. Rudner, David Rügamer, Yee Whye Teh, Max Welling, Andrew Gordon Wilson, and Ruqi Zhang. Position: Bayesian deep learning is needed in the age of large-scale AI. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 39556–39586, 2024.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, 2014.

Hippolyt Ritter, Aleksandar Botev, and David Barber. A scalable Laplace approximation for neural networks. In *International Conference on Learning Representations*, 2018.

Tim G. J. Rudner, Zonghao Chen, Yee Whye Teh, and Yarin Gal. Tractable function-space variational inference in Bayesian neural networks. In *Advances in Neural Information Processing Systems*, 2022.

T. Salimans and D. P. Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In *Advances in Neural Information Processing Systems*, 2016.

Mrinank Sharma, Sebastian Farquhar, Eric Nalisnick, and Tom Rainforth. Do Bayesian neural networks need to be fully stochastic? In *International Conference on Artificial Intelligence and Statistics*, 2023.

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 *International Conference on Machine Learning*, 2024.

Emanuel Sommer, Lisa Wimmer, Theodore Papamarkou, Ludwig Bothmann, Bernd Bischl, and David Rügamer. Connecting the dots: Is mode-connectedness the key to feasible sample-based inference in Bayesian neural networks? In *International Conference on Machine Learning*, 2024.

Suvrit Sra. Directional statistics in machine learning: A brief review. *arXiv preprint arXiv:1605.00316*, 2016.

Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014.

Shengyang Sun, Changyou Chen, and Lawrence Carin. Learning structured weight uncertainty in Bayesian neural networks. In *International Conference on Artificial Intelligence and Statistics*, 2017.

Shengyang Sun, Guodong Zhang, Jiaxin Shi, and Roger Grosse. Functional variational Bayesian neural networks. In *International Conference on Learning Representations*, 2019.

Brian Trippe and Richard Turner. Overpruning in variational Bayesian neural networks. *arXiv preprint arXiv:1801.06230*, 2018.

Aki Vehtari, Andrew Gelman, Daniel Simpson, Bob Carpenter, and Paul-Christian Bürkner. Rank-normalization, folding, and localization: An improved $\hat{R}$ for assessing convergence of MCMC. *Bayesian Analysis*, 16(2), 2021.

Ruosi Wan, Zhanxing Zhu, Xiangyu Zhang, and Jian Sun. Spherical motion dynamics: Learning dynamics of normalized neural network using SGD and weight decay. In *Advances in Neural Information Processing Systems*, 2021.

Sida Wang and Christopher Manning. Fast dropout training. In *International Conference on Machine Learning*, 2013.

Florian Wenzel, Kevin Roth, Bastiaan S. Veeling, Jakub Swiatkowski, Linh Tran, Stephan Mandt, Jasper Snoek, Tim Salimans, Rodolphe Jenatton, and Sebastian Nowozin. How good is the Bayes posterior in deep neural networks really? In *International Conference on Machine Learning*, pp. 10248–10259, 2020.

Andrew Gordon Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. In *Advances in Neural Information Processing Systems*, 2020.

Andrew T. A. Wood. Simulation of the von Mises Fisher distribution. *Communications in Statistics - Simulation and Computation*, 23(1):157–164, 1994. doi: 10.1080/03610919408813161.

Anqi Wu, Sebastian Nowozin, Edward Meeds, Richard E. Turner, José Miguel Hernández-Lobato, and Alexander L. Gaunt. Deterministic variational inference for robust Bayesian neural networks. In *International Conference on Learning Representations*, 2019.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Adam X. Yang, Maxime Robeyns, Xi Wang, and Laurence Aitchison. Bayesian low-rank adaptation for large language models. In *International Conference on Learning Representations*, 2024.

A Moments of the pre-activation under a von Mises–Fisher posterior

For $w \sim \mathrm{vMF}(\mu, \kappa)$ on S^{D-1}, $\mathbb{E}[w] = A_D(\kappa)\,\mu$ with $A_D(\kappa) = I_{D/2}(\kappa)/I_{D/2-1}(\kappa)$, and (Mardia & Jupp, 2000; Kurz & Hanebeck, 2015)

$$\begin{aligned} \mathrm{Cov}(w) &= C_{\parallel}(\kappa)\,\mu\mu^{\top} + C_{\perp}(\kappa)\,(I - \mu\mu^{\top}), \\ C_{\perp} = \frac{A_D(\kappa)}{\kappa}, &\qquad C_{\parallel} = 1 - \frac{(D-1)A_D(\kappa)}{\kappa} - A_D(\kappa)^2, \end{aligned} \tag{4}$$

so that $C_{\parallel} = A_D'(\kappa)$ and $C_{\parallel} + (D-1)C_{\perp} = 1 - A_D(\kappa)^2$. The asymptotics are $A_D(\kappa) \approx 1-(D-1)/2\kappa$ for large κ and $A_D(\kappa) \approx \kappa/D$ for small κ, the latter from $I_\nu(z) \approx (z/2)^\nu/\Gamma(\nu+1)$. The pre-activation variance is $\sigma_u^2 = C_{\parallel}(\mu^{\top}x)^2 + C_{\perp}(\|x\|^2 - (\mu^{\top}x)^2)$. With $(\mu^{\top}x)^2 \approx \|x\|^2/D$ for a normalised input, the trace identity gives $\sigma_u^2 = \|x\|^2\,(1 - A_D(\kappa)^2)/D$. Its large-$\kappa$ expansion, $(\|x\|^2/\kappa)(1 - 1/D)(1 - (D-2)/2\kappa)$, turns over below $\kappa \approx D$; the implementation uses

$$\sigma_u^2 = \frac{\|x\|^2}{\kappa + D}, \tag{5}$$

which has the right limits at both ends ($\|x\|^2/\kappa$ for $\kappa \gg D$; $\|x\|^2/D$, the variance under the uniform prior, for $\kappa \to 0$), is monotone throughout, and lies below the exact form by at most 20%, near $\kappa/D \approx 0.7$, and by 4% at $\kappa/D = 10$. Because A_D scales signal and noise together, dropping it changes only the layer gain, which a downstream normalisation removes; the κ-form $u = A_D\,[(x \cdot \hat{\mu}) + \sigma_{\text{eff}}\,\mathrm{rms}(x)\,\varepsilon] + b$ and the σ-form of Eq. (2) give the same noise-to-signal ratio and KL to within 0.2% over $\kappa = 10$ to 10^6.

B THE KL DIVERGENCE AND ITS CLOSED FORM

With $C_D(\kappa) = \kappa^{D/2-1}/((2\pi)^{D/2} I_{D/2-1}(\kappa))$ and $|S^{D-1}| = 2\pi^{D/2}/\Gamma(D/2)$,

$$\mathrm{KL}(q \,\|\, \mathrm{unif}) = \kappa A_D(\kappa) + \log C_D(\kappa) + \log |S^{D-1}|. \tag{6}$$

For large κ, $\mathrm{KL} \approx \frac{D-1}{2} \log \frac{\kappa}{2\pi e} + \log |S^{D-1}|$ (Sra, 2016; Banerjee et al., 2005); for small κ, $\mathrm{KL} \approx \kappa^2/2D$. By Stirling's formula the $\log D$ terms of the large-κ form cancel against $\log |S^{D-1}|$, leaving $\mathrm{KL} \approx \frac{D-1}{2} \log(\kappa/D) + O(1)$. In the σ_{eff} parameterisation, $\kappa \approx D/\sigma_{\mathrm{eff}}^2$ at small σ_{eff} and $\kappa \approx D/\sigma_{\mathrm{eff}}$ at large σ_{eff}, giving $\mathrm{KL} \approx -(D-1)\log \sigma_{\mathrm{eff}}$ and $\mathrm{KL} \approx \frac{D}{2}\sigma_{\mathrm{eff}}^{-2}$ respectively. Eq. (3) interpolates the two with the correct prefactors at both ends. Its relative error against the exact expression is controlled by κ/D: 19.1% at $\kappa/D = 0.3$, 17.9% at 1, 1.73% at 10, 0.09% at 100, 0.005% at 1000, with every D on the same curve in the high-error region. At the operating point of the Fashion-MNIST runs ($D = 512$, $\kappa = 388$) the closed form overestimates the exact KL by 20.3%; the corresponding figures for VD, MFVI and the HMC concentration are 20.2%, 17.6% and 21.7%.

The dictionary. From $\sigma_u^2 = D/(\kappa + D)$ for a normalised input and $\sigma_{\mathrm{eff}} = \sigma_u / A_D(\kappa)$,

$$\sigma_{\mathrm{eff}} = \frac{\sqrt{D/(\kappa + D)}}{A_D(\kappa)}, \qquad \mathbb{E}_q \cos \angle(w, \mu) = A_D(\kappa). \tag{7}$$

For large D, $A_D(\kappa)$ is a function of κ/D in both limits (κ/D for small, $1 - 1/(2\kappa/D)$ for large), and so are σ_{eff} and $\mathrm{KL}/(D-1)$, which is the invariance used in Section 2; across $D = 256$, 512 and 4096 the angle at fixed σ_{eff} agrees to within 0.2% and $\mathrm{KL}/(D-1)$ to within 0.4%, and down to $D = 27$ the angle stays within 0.4°. Table 4 gives reference points.

Table 4: The dictionary at $D = 512$: learned noise scale, concentration per dimension and the angle between a draw and the mean direction.

σ_{eff}	0.10	0.33	1.0	3.0	10	20
κ/D	100	9.2	1.22	0.32	0.096	0.049
angle	5.7°	18.7°	47.8°	73.1°	84.5°	87.2°

C WITHIN-UNIT ANISOTROPY: PROOF OF PROPOSITION 2

Let the log posterior near its mode be $-\frac{1}{2} t^\top H t$ in tangent coordinates $t \in \mathbb{R}^n$, $n = D - 1$. A von Mises–Fisher with concentration κ has tangent covariance $(1/\kappa) I$ to leading order. The expected log-likelihood under q contributes $-\frac{1}{2}\,\mathrm{tr}(H)/\kappa$ to the ELBO and the KL contributes $-\frac{n}{2}\log\kappa$ up to constants (large-κ form); the stationary point in κ is $\kappa_q = \mathrm{tr}(H)/n$, the arithmetic mean of the eigenvalues. The true angular spread is $\mathbb{E}\|t\|^2 = \mathrm{tr}(H^{-1})$, matched by a von Mises–Fisher with $\kappa_{\mathrm{true}} = n/\mathrm{tr}(H^{-1})$, the harmonic mean. The ratio $\mathrm{tr}(H)\mathrm{tr}(H^{-1})/n^2 \geq 1$ is Cauchy–Schwarz applied to $(\sqrt{\lambda_i})$ and $(1/\sqrt{\lambda_i})$, with equality iff all λ_i are equal.

D THE WIDTH LAW

A layer of M units with per-unit noise σ_{eff} injects independent noise into each unit while the signal is coherent across units, so the layer-level signal-to-noise ratio is $\sqrt{M}$ times the per-unit ratio. The effective regularisation a layer exerts on the network is set by its layer-level ratio. Holding that fixed as M grows therefore requires $\sigma_{\mathrm{eff}}^* \propto M^{1/2}$. A sweep in which fan-in grows with width ($[784, M, M, M]$) does not test the law, because the KL then scales as MD; the fitted exponent is $+0.919$; Section 6 uses this sweep for the high-noise regime.

E THE CORRELATED VARIANT

Give q the cross-unit covariance $\Sigma = D_\sigma R D_\sigma \otimes I_{D-1}$ with R unit-diagonal. Then $\Sigma_{kk} = \sigma_k^2$, so σ_{eff} remains the marginal width, and

$$\begin{aligned} \mathrm{KL} &= \sum_k \mathrm{KL}_{\mathrm{CAP}}(\sigma_{\text{eff},k}, D) - \frac{D-1}{2} \log\det R, \\ \log\det R &= \log\big(1 + (K-1)\rho\big) + (K-1)\log(1-\rho) \end{aligned} \tag{8}$$

for the equicorrelated case. Since $\log\det R \le 0$, correlation costs KL. Sampling is $\varepsilon = Lz$ with $L = \mathrm{chol}(R)$, one $K \times K$ Cholesky per step. Learning ρ takes the erf teacher's overconfidence from 8.35 to 1.29. Fixing ρ at the measured 0.87 gives 66.3; where correlation is small, learning ρ leaves the fit unchanged (1.89 to 1.88).

F THE ℓ_1 PRIOR ON THE SPHERE

With $p(w) \propto \exp(-\lambda\|w\|_1)$ on S^{D-1} the KL gains $\lambda\,\mathbb{E}_q\|w\|_1 + \log Z(\lambda, D)$; $\log Z$ is constant in the variational parameters and enters no gradient. Since q concentrates on μ, $\mathbb{E}_q\|w\|_1 = \|\mu\|_1 + O(1/\kappa)$. For w uniform on S^{D-1}, $\mathbb{E}\|w\|_1 = D\,\Gamma(D/2)/(\sqrt{\pi}\,\Gamma((D+1)/2)) \to \sqrt{2D/\pi}$; the reported ℓ_1 ratio is $\|\mu\|_1/\sqrt{2D/\pi}$, with floor $1/\sqrt{2D/\pi}$ for a single axis (0.078 at $D = 256$). Because $\|w\|_2 = 1$, the ℓ_1 norm of a weight direction is a bounded, scale-free measure of how many inputs a unit reads, from 1 on one axis to $\sqrt{D}$ when dense. Table 5 sweeps λ on Fashion-MNIST at width 256, $\beta = 0.3$, two seeds. At $\lambda = 3$ the direction is about half as dense as random, at an accuracy within seed noise of the unpenalised model, from a prior change on the sphere alone; on the same statistic this is comparable to what variational dropout's per-weight prior reaches (VD's own pruning mask, 0.895, is a magnitude statistic with no CAP counterpart). The largest λ, 30, reaches 0.123 at a cost of 1.4 accuracy points. This is fan-in sparsity, which inputs a unit reads. Removing whole units requires a magnitude, which is why the structured pruning of Oh et al. (2020) lives on their radial posterior.

Table 5: Fan-in sparsity from an ℓ_1 prior on the sphere. Fashion-MNIST, width 256, $\beta = 0.3$, two seeds.

λ	ℓ_1 ratio	fraction below $0.1/\sqrt{D}$	test acc.
0	0.997	0.07	0.866
1	0.694	0.41	0.863
3	0.438	0.72	0.863
10	0.222	0.92	0.856
VD, per weight	0.490	0.55	0.860

G THE TOY SETUP AND THE GATE FOR THE POSTERIOR-FIT COMPARISON

Toy setup. Throughout the teacher–student experiments, $x \sim \mathcal{N}(0, I_D)$ and $y \sim \mathrm{Bernoulli}\big(\sigma(\lambda \sum_{k=1}^K a_k\,\phi(x^\top B_k + b))\big)$ with $\phi = \mathrm{relu}$ ($\phi = \mathrm{erf}$ for the erf teachers of Section 3), unit-norm teacher rows B_k, $a_k = 1/\sqrt{K}$ and λ either fixed or set so that the teacher's Bayes accuracy takes a target value. "Angle" always means the mean angle between a posterior draw and the posterior mean direction; 90° is the uniform prior. Angular errors to the teacher use a Hungarian match between student and teacher rows, since the units are permutation-symmetric.

Selection of settings. Table 6 gives the full sweep behind Section 4. The two settings meeting both criteria are those at $N = 8000$ with bias -1.0 and -0.5.

Table 6: HMC sweep. W_1 is the 1-Wasserstein distance in degrees between CAP's angular law and HMC's, averaged over units.

b	N	coin-flip labels	$\hat{R}$	ESS	HMC angle	W_1 CAP
-2.0	2000	89.7%	1.010	921	$78.6°$	4.51
-2.0	8000	88.6%	1.514	37	$69.2°$	9.53
-1.0	2000	39.6%	1.240	39	$67.2°$	10.38
-1.0	8000	38.6%	1.002	3569	$12.9°$	2.18
-0.5	2000	13.0%	1.217	60	$68.9°$	4.20
-0.5	8000	12.7%	1.001	2849	$14.6°$	3.05

H UCI REGRESSION

Architecture and splits follow Hernández-Lobato & Adams (2015) and Gal & Ghahramani (2016) (20 splits; 5 for protein), with 50-unit hidden layers, learned observation precision τ following Oh et al. (2020) and hyperparameters selected on a validation split carved from training. All arms share one implementation; absolute numbers depend on the training budget and are not compared to published values. Table 7 gives the four datasets on which every arm was tuned: CAP takes the best test log-likelihood on two and the best RMSE on all four, and the learned observation precision lands within a few percent across CAP, MFVI and VD, so it does not drive the differences. All eight datasets independently selected $\beta \in [0.02, 0.04]$, an order of magnitude below the classification value of Section 6. Protocol: $\log \tau$ has its own learning rate; the first layer is stochastic, since a deterministic one makes a one-hidden-layer network fully deterministic; the output head is unnormalised in the no-normalisation arms; and the epoch budget is fixed per dataset.

Table 7: UCI regression, every arm tuned on validation under the same code. Test log-likelihood and RMSE, mean over the standard splits.

dataset	CAP LL	best other LL	CAP RMSE	best other RMSE
yacht	-2.433	-2.441 (MFVI)	1.694	1.782 (VD)
Boston	-2.548	-2.555 (MFVI)	3.009	4.424 (MFVI)
energy	-1.870	-1.852 (MFVI)	1.172	1.484 (MC dropout)
concrete	-3.323	-3.281 (MFVI)	6.305	8.034 (MFVI)

I IMPLEMENTATION AND TRAINING PROTOCOL

Layers. Weight-normalised linear layers with fixed unit gain, followed by batch normalisation without affine parameters; biases deterministic. CAP noise is injected at the pre-activation as in Eq. (2), with $\log \sigma_{\text{eff}}$ learned per unit or per layer under its own learning rate. Baselines: mean-field Gaussian (per-weight μ, σ), variational dropout (Molchanov et al., 2017), radial (Farquhar et al., 2020), MC dropout with a single rate and a deterministic control. All arms share the architecture, optimiser and data pipeline.

Matched initialisation. All learned-noise arms start at the same realised noise-to-signal ratio, measured from the forward pass, since their own parameterisations would otherwise start them up to $400\times$ apart (CAP 1.14, MFVI 0.26, VD 0.12, radial 0.003); one noise-free epoch precedes the variational objective. Cross-method noise levels are reported as the realised ratio, which is the only quantity on a common scale; the per-method σ parameters normalise differently (by row norm, by matrix Frobenius norm, or not at all).

Fashion-MNIST setup. Fashion-MNIST (Xiao et al., 2017), per-feature standardised inputs; architecture $784 \to 256 \to M \to 256 \to 10$ with batch normalisation and ReLU; 50 epochs, batch

256, Adam at 10^{-3} with the noise parameters at 0.03; the first block deterministic unless stated, blocks 1 and 2 carrying the noise, and a deterministic head.

Evaluation. $S = 16$ stochastic passes unless stated; ECE (Naeini et al., 2015; Guo et al., 2017) with 15 bins; the mean-network evaluation re-estimates batch-norm statistics with the noise off by resetting the running statistics and accumulating over two passes of the training set (Table 8; at width 1024, where $\sigma_{\text{eff}} \approx 18$, the stale statistics give 0.68; MC dropout needs no re-estimation because its noise is small enough for the two regimes to coincide; NLL does not recover for any stochastic arm, since a single pass carries no predictive uncertainty).

Table 8: Test accuracy of the mean network, Fashion-MNIST, width 1024.

arm	MC, $S = 16$	mean, stale BN	mean, BN re-estimated
CAP, per layer	0.866	0.683	0.861
CAP, per unit	0.864	0.677	0.868
MFVI	0.866	0.335	0.588
MC dropout	0.871	0.870	0.873

HMC on Fashion-MNIST samples the directions of the stochastic layer under the uniform prior with the rest of the network fixed, with four chains and split-$\hat{R}$ and ESS reported (Vehtari et al., 2021); von Mises–Fisher draws use Wood (1994) and the reparameterisation follows Kingma & Welling (2014) and Rezende et al. (2014); conversion from σ_{eff} to κ is checked against sample-and-invert.

J Additional figures and the width axis

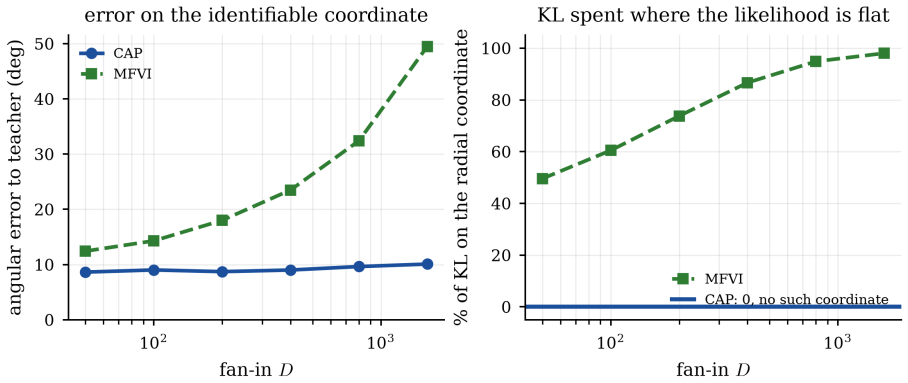

Figure 5: What a factorised Gaussian spends on the radial coordinate. Toy setup with a batch-normalised pre-activation, so the posterior's radial marginal is exactly the prior's; D from 50 to 1600 with $N = 67KD$, so labels per parameter are held fixed. Left: angular error to the teacher against fan-in, after a Hungarian match of student to teacher units; CAP goes from $8.64°$ to $10.09°$, MFVI from $12.43°$ to $49.49°$. Right: the share of MFVI's KL that falls on the radial direction, computed in closed form by splitting $\mathrm{KL}(\mathcal{N}(\mu, \sigma^2 I) \,\|\, \mathcal{N}(0, I/D))$ along $\hat{\mu}$ and the $D - 1$ tangent directions; it rises from 49.6% to 98.0%. CAP's share is zero because the coordinate does not exist in the family.

The width axis for overconfidence. Table 9 gives the sweep behind Section 5: $D = 200$, $rN/D = 640$, K from 4 to 64. Overconfidence falls monotonically to $K = 32$; at $K = 64$ ($K/D = 0.32$) the realised angle is $66°$ and the problem is past the recoverable boundary, where more data made the overconfidence worse (10.4 at $rN/D = 320$ to 19.7 at 1280), the signature of a direction the data no longer determine. The cross-unit correlation was flat at 0.158 to 0.161 in every configuration.

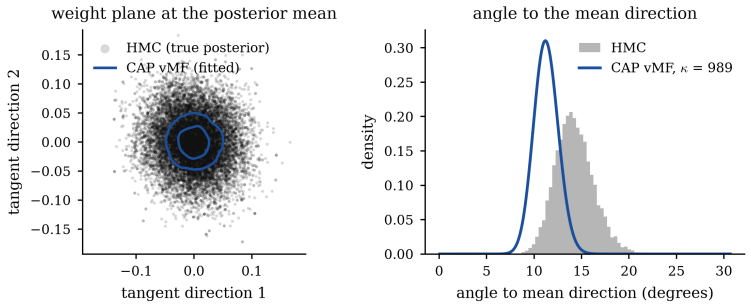

Figure 6: CAP against HMC at a selected setting: $D = 40$, $K = 6$, $N = 8000$, ReLU teacher with bias -0.5; HMC over the directions under the uniform prior, four chains, 3000 warm-up and 4000 samples, $\hat{R} = 1.002$, ESS 2651. The angle to the mean direction: HMC's histogram sits at $14.25°$ and CAP's von Mises–Fisher at $\kappa = 989$ sits on it, $3.05°$ away in 1-Wasserstein distance and $1.5\times$ overconfident in κ, the anisotropy of Proposition 2 in a setting where the sampler converged. The other selected setting gives $2.18°$.

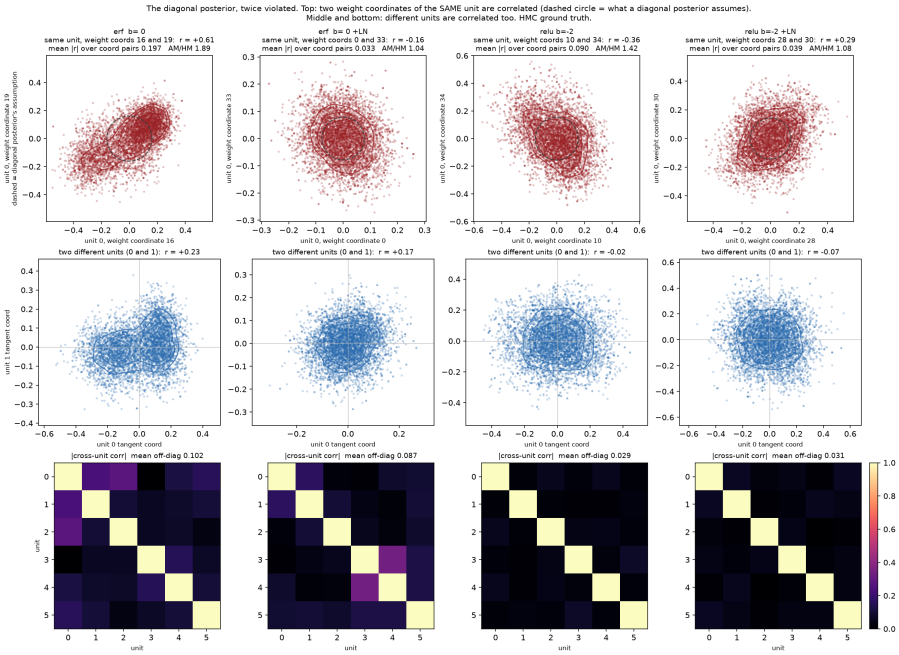

Figure 7: The two factorisation assumptions in all four conditions of Table 1. Rows: two coordinates of one unit; a tangent coordinate of unit 0 against one of unit 1; the full cross-unit correlation matrix. Columns: erf with zero bias, the same with layer normalisation, ReLU with bias -2, the same with layer normalisation. Layer normalisation removes most of the within-unit structure and leaves the cross-unit structure where it was; sparse activation removes most of the cross-unit structure and part of the within-unit structure.

Table 9: Overconfidence against the number of units at $D = 200$, $rN/D = 640$.

K	max overlap	ρ	realised angle	predicted angle	overconfidence
4	0.075	0.158	$2.92°$	$1.87°$	2.454
8	0.126	0.159	$4.05°$	$2.74°$	2.183
16	0.168	0.159	$6.23°$	$4.64°$	1.805
32	0.257	0.158	$11.29°$	$8.73°$	1.678
64	0.257	0.159	$66.07°$	$21.74°$	13.908

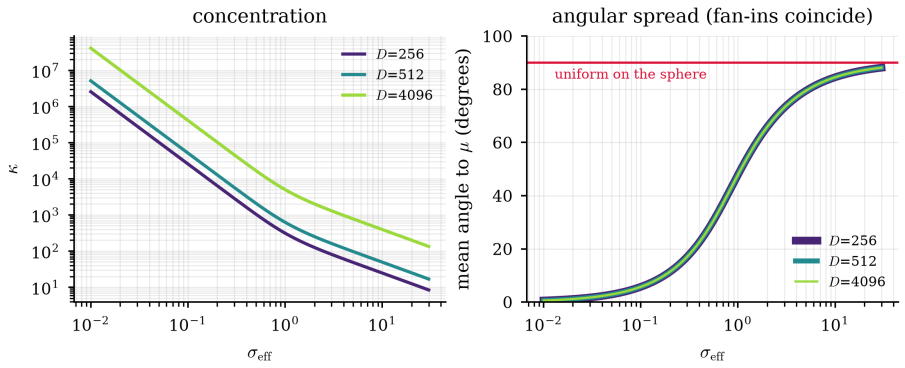

Figure 8: The dictionary across fan-in. Both panels are analytic, against σ_{eff}, for $D = 256, 512, 4096$. Left: κ; the curves are separated, so a value of κ means nothing without the fan-in. Right: the mean angle between a draw and the mean direction; the three curves coincide to the width of the line (drawn at decreasing widths so all three show). The red line is 90°, the uniform prior.

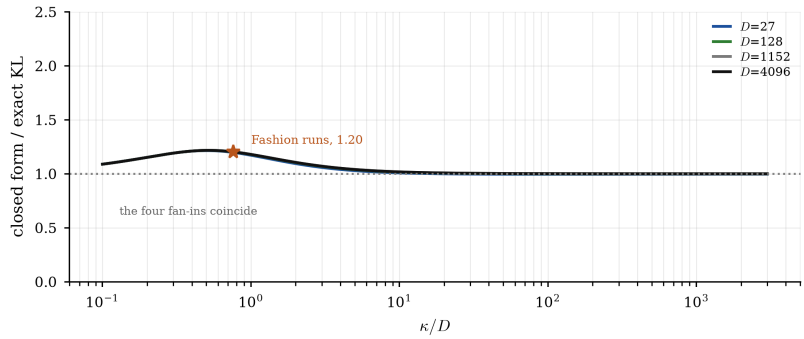

Figure 9: The closed-form KL of Eq. 3 divided by the exact von Mises–Fisher divergence, against κ/D. The ratio is one over most of the range and the four fan-ins coincide, so the accuracy depends on κ/D alone. It peaks at 1.22 near $\kappa/D = 0.5$; the star marks the operating point of the Fashion-MNIST networks of Section 4 (fan-in 512, learned $\kappa = 388$), where the closed form overestimates the divergence by 20%. Since the overestimate is close to constant across the range a network visits, it acts as a rescaling of β.

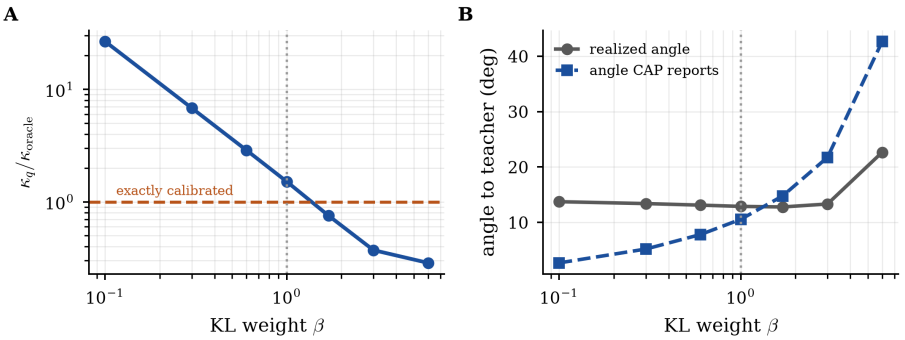

Figure 10: The KL weight against a known teacher. At $N = 20$k, $M = 1024$, the noise-free warm-up epoch moves NLL by about 2%, while putting CAP on the raw-pixel input block moves it by 29% and drags the optimum to $\beta = 0.1$; ZCA-whitening the inputs leaves the optimum at $\beta = 1$. With a mean-field Gaussian on the readout as well, likelihood and calibration are both best at exactly $\beta = 1$. $D = 40$, $K = 6$, $N = 8000$, ReLU teacher with bias -1, two seeds. (A) Overconfidence $\kappa_q/\kappa_{\text{oracle}}$ against β on log-log axes with 1 marked; it crosses near $\beta \approx 1.2$ and reaches 26.7× at $\beta = 0.1$. (B) The realised angle to the teacher, flat at about 13°, against the angle CAP reports, which goes from 2.7° to 21.8°.

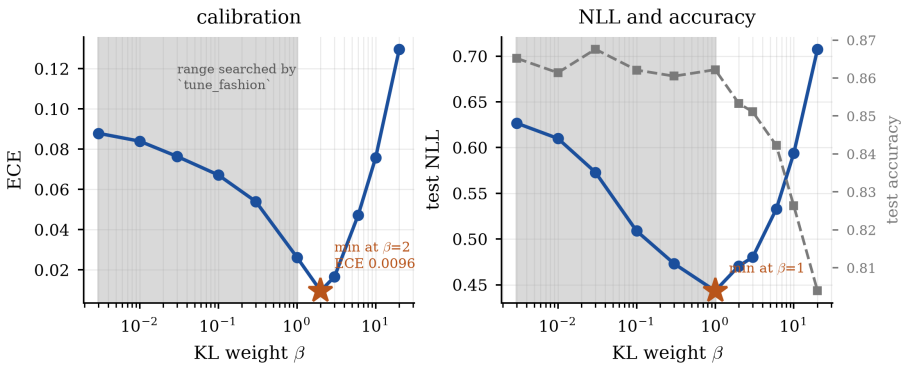

Figure 11: ECE and NLL against β for CAP at $N = 20{,}000$, $M = 1024$, β from 0.003 to 20, with the range searched by the tuning run shaded. The interior minima (ECE at $\beta = 2$, NLL at $\beta = 1$) sit above the searched range; accuracy falls from 0.862 to 0.804 as β goes from 1 to 20.

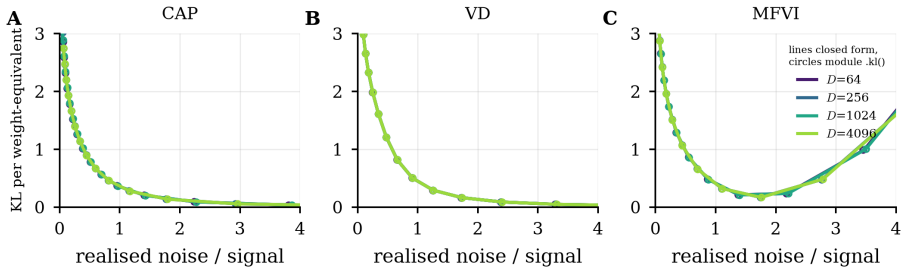

Figure 12: KL per weight-equivalent against the noise actually injected, for CAP, mean-field and variational dropout with several widths overlaid. Lines are the closed forms, circles the modules' own KL. A log-y version is in the supplementary files.

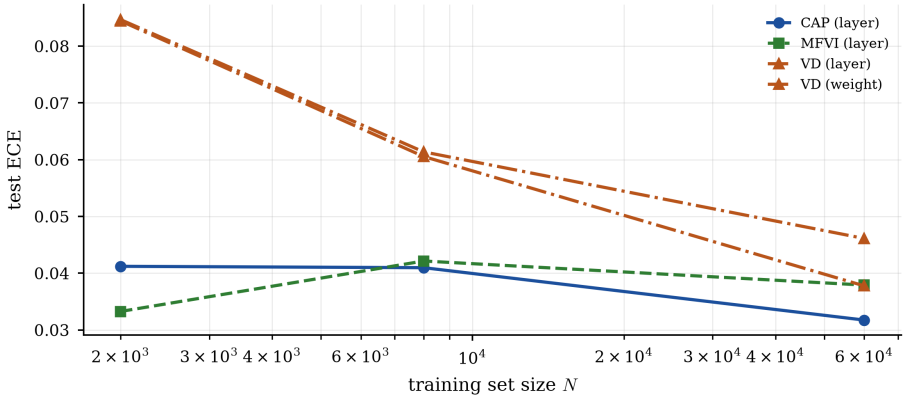

Figure 13: Calibration against training-set size on Fashion-MNIST, one noise scalar per layer. Variational dropout is badly calibrated on small data and recovers only as N grows: ECE 0.084 at $N = 2000$ against 0.038 at $N = 60{,}000$. CAP is flat and then improves, 0.041 to 0.032, and the mean-field Gaussian is flat at 0.033 to 0.038.

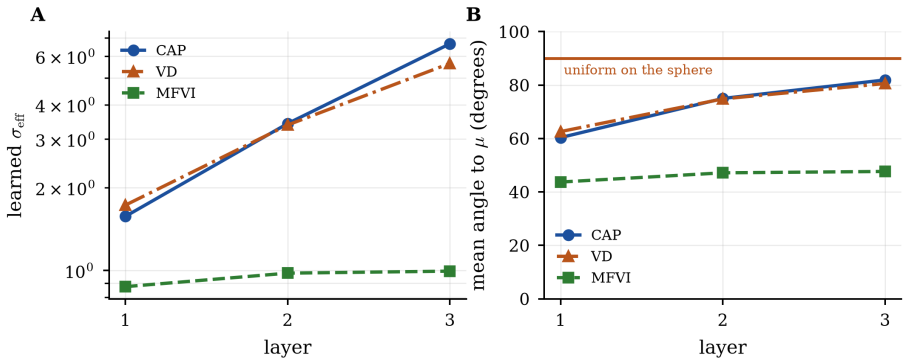

Figure 14: Learned noise scale by depth in a CIFAR-10 CNN, read on the angular scale of Section 2. Fan-in grows with depth $(288, 576, 1152)$ and so does the directional uncertainty: CAP goes $\sigma_{\text{eff}} = 1.57, 3.42, 6.66$, that is $60°, 75°, 82°$, and variational dropout tracks it at $1.73, 3.38, 5.65$ $(63°, 75°, 81°)$. The mean-field Gaussian is flat at $0.87, 0.98, 0.99$ $(44°$ to $48°)$, so it does not adapt to fan-in at all.