Scalable Mean-Field Variational Inference through Deterministic Teacher-Guided Knowledge Distillation

Anonymous authors
Paper under double-blind review

Abstract

Scaling mean-field variational inference (MFVI) to large-scale deep networks remains challenging: in high dimensions, Gaussian posteriors concentrate probability mass on thin hyperspherical shells far from the mean (the "soap-bubble" phenomenon), so sampled weights produce high-variance gradient estimates that destabilize ELBO training. On ImageNet, vanilla MFVI achieves only 21.57% top-1 accuracy with a ResNet-18 backbone, indicating severe optimization degradation relative to its deterministic counterpart. We introduce a decoupled training framework that addresses this issue at the level of the optimization objective. A pretrained teacher network supervises the mean parameters μ through an output-space knowledge distillation term, while the variance parameters ρ are optimized exclusively through the standard ELBO. The combined objective admits a hybrid MAP-VI interpretation in which the distillation term defines a teacher-induced pseudo-prior on μ, recovering standard MFVI as a special case. Under standard local regularity conditions, we prove a dimension-independent confinement bound showing that the optimum lies within a controllable neighborhood of the teacher's effective solution. Empirically, the method reaches 71.58% top-1 accuracy and ECE 0.0251 on ImageNet with ResNet-18, and is complementary to existing techniques such as MOPED and Radial reparameterization.

1 Introduction

Deep neural networks have achieved remarkable success across diverse domains, yet their tendency to produce overconfident predictions—even on out-of-distribution samples—poses significant challenges for safety-critical applications such as medical diagnosis, autonomous driving, and scientific discovery. Bayesian Neural Networks (BNNs) offer a principled solution by placing probability distributions over network weights, thereby modeling epistemic uncertainty and providing a mathematically grounded framework for uncertainty quantification (Kendall & Gal, 2017; Blundell et al., 2015). However, despite theoretical appeal and methodological advances in stochastic variational inference (Blundell et al., 2015; Graves, 2011), dropout-based approximations (Gal & Ghahramani, 2016), and modern reparameterization techniques (Farquhar et al., 2020), BNNs remain computationally prohibitive and memory-intensive, severely limiting their applicability to large-scale datasets such as ImageNet.

Recent efforts to address these scalability challenges have focused on two complementary strategies: constraining posterior complexity and stabilizing optimization in high-dimensional parameter spaces. Radial BNNs, for instance, mitigate the "soap-bubble" pathology—where weight distributions collapse onto a thin shell in parameter space—by adopting a radial reparameterization that maintains meaningful variance throughout training (Farquhar et al., 2020). The MOPED framework takes an alternative approach, leveraging informed priors that constrain the range of variational parameters relative to pretrained deterministic weights, thereby anchoring posterior distributions and preventing variance explosion (Krishnan et al., 2020). While these methods successfully stabilize training, they impose structural constraints that may inadvertently limit the network's capacity to capture rich, multimodal uncertainty—a critical limitation when dealing with complex, high-dimensional data distributions characteristic of large-scale visual recognition tasks.

In parallel, knowledge distillation (KD) has emerged as a powerful paradigm for transferring the learned representations of high-capacity teacher networks to more compact student models (Hinton et al., 2015). By matching student predictions to the softened outputs of a teacher, distillation enables significant model compression while preserving much of the teacher's representational power (Hinton et al., 2015). Despite its widespread adoption in deterministic settings, the integration of knowledge distillation with Bayesian neural networks remains largely unexplored. Naive applications—where distillation losses are simply added to the standard Bayesian objective—often fail to reconcile the conflicting demands of deterministic teacher guidance and the Bayesian imperative for flexible posterior inference, leading to suboptimal uncertainty estimates or convergence difficulties.

Our Contributions. In this work, we propose a novel *decoupled distillation framework* that combines knowledge distillation with scalable Bayesian inference, explicitly addressing both computational tractability and robust uncertainty quantification. Our key insight is to decompose the optimization of variational parameters in a mean-field BNN into two independent but complementary objectives, operating on distinct parameter subsets.

Specifically, we denote the variational parameters as $\boldsymbol{\theta} = \{\boldsymbol{\mu}, \boldsymbol{\rho}\}$, where $\boldsymbol{\mu}$ represents the mean of the weight distributions and $\boldsymbol{\rho}$ parameterizes the variance through $\boldsymbol{\sigma} = \log(1+\exp(\boldsymbol{\rho}))$ using the softplus transformation. Our contributions can be summarized as follows:

- **Decoupled training scheme.** We propose a training scheme that separates representation learning from uncertainty calibration: the mean parameters $\boldsymbol{\mu}$ are guided by an output-space distillation term $\mathcal{L}_{\text{KD}}(\boldsymbol{\mu})$, while the variance parameters $\boldsymbol{\rho}$ are optimized exclusively through the ELBO. The combined objective preserves the original Bayesian prior on weights and adds the distillation term only as a function of $\boldsymbol{\mu}$, leaving the variance free to be calibrated by the data through standard variational inference.
- **Probabilistic interpretation.** We show that the combined objective admits a hybrid MAP-VI reading in which the distillation term induces a teacher-derived pseudo-prior on the variational mean, $p_T(\boldsymbol{\mu}) \propto \exp(-\alpha\, G(\boldsymbol{\mu}))$. This perspective recovers standard MFVI as a special case ($\alpha \to 0$) and distinguishes the method from MOPED-style informed weight-space priors, which act on $p(\theta)$ rather than on $\boldsymbol{\mu}$.
- **Confinement bound (Theorem 3.1).** Under standard local regularity conditions, we prove that any stationary point $\boldsymbol{\mu}^*$ of the combined objective satisfies a dimension-independent bound $\|\boldsymbol{\mu}^* - \boldsymbol{\mu}^T\| \leq (1+\alpha)C/(\alpha\lambda - L)$, which is monotonically decreasing in α. The bound provides a geometric justification for the empirical stability of the proposed training scheme.
- **Empirical demonstration at scale.** We evaluate the method on CIFAR-100 and ImageNet-1K using ResNet architectures. On ImageNet, where vanilla MFVI degrades severely (21.57%), our method reaches 71.58% top-1 accuracy with strong calibration (ECE 0.0251). The method is also complementary to existing techniques such as MOPED and Radial reparameterization.

The decoupled formulation has two practical advantages. First, separating deterministic supervision (mean alignment) from stochastic inference (variance learning) simplifies the optimization landscape: the variance is no longer pulled toward matching deterministic teacher outputs, which would otherwise compress epistemic uncertainty estimates. Second, the distillation term $G(\boldsymbol{\mu})$ requires only a single deterministic forward pass per batch, adding negligible overhead to standard BNN training. The remainder of the paper develops these contributions in detail, presents the confinement result and its assumptions, and validates the method experimentally.

2 Related Work

Our work lies at the intersection of scalable Bayesian deep learning and knowledge distillation. We organize our discussion around three central themes: approximate inference methods for BNNs, approaches to mitigate

pathologies in high-dimensional variational inference, and the emerging integration of transfer learning with Bayesian neural networks.

2.1 Scalable Approximate Inference for Bayesian Neural Networks

The fundamental challenge in Bayesian deep learning is computing or approximating the intractable posterior distribution over network weights. Early work by MacKay (1992) and Neal (2012) established the theoretical foundations but relied on computationally expensive Markov Chain Monte Carlo (MCMC) methods that scale poorly to modern deep architectures. This scalability barrier has motivated the development of several approximate inference paradigms.

Sampling-based approximations. Monte Carlo Dropout (Gal & Ghahramani, 2016) reinterprets standard dropout as approximate variational inference in deep Gaussian processes, enabling uncertainty quantification through multiple stochastic forward passes at test time. Deep Ensembles (Lakshminarayanan et al., 2017) adopt an even simpler approach: training multiple independent networks with different random initializations and treating their predictions as samples from an implicit posterior. While both methods are remarkably effective in practice and computationally efficient, they suffer from a fundamental limitation: they assign exactly zero probability to almost all regions of weight space, yielding posteriors with discrete support. This discrete support property makes these approaches theoretically unsuitable for sequential Bayesian updating (Farquhar et al., 2020) and conceptually inconsistent with the Bayesian paradigm, which assumes smooth, continuous distributions over parameters. Recent work has questioned whether these methods truly perform Bayesian inference or merely provide useful heuristics for uncertainty estimation (Foong et al., 2019).

Mean-field variational inference. A more principled alternative is variational inference (VI), which approximates the true posterior with a tractable parametric distribution by minimizing the Kullback-Leibler divergence (Jordan et al., 1999). The seminal work of Graves (2011) demonstrated practical VI for neural networks, while Blundell et al. (2015) introduced Bayes by Backprop, a scalable gradient-based algorithm that employs the reparameterization trick to enable end-to-end training. These mean-field variational inference (MFVI) approaches assume a fully factorized Gaussian posterior, doubling the parameter count but maintaining computational tractability.

However, MFVI exhibits severe pathologies in high-dimensional spaces. As Farquhar et al. (2020) observed, multivariate Gaussians in high dimensions concentrate their probability mass on a thin hyperspherical shell far from the mean—the so-called "soap-bubble" phenomenon. During training, weight samples drawn from this shell have highly variable magnitudes, leading to high-variance gradient estimates that destabilize optimization. This problem intensifies with network depth and width, precisely when uncertainty quantification becomes most valuable. Moreover, MFVI's factorial approximation cannot capture posterior correlations between weights, limiting its expressiveness even when optimization succeeds.

2.2 Addressing Pathologies in Variational Bayesian Deep Learning

Several recent works have proposed remedies to the limitations of standard MFVI, each making different trade-offs between expressiveness, computational cost, and ease of optimization.

Structured posteriors. One direction relaxes the mean-field assumption to capture richer posterior structures. Approaches such as matrix-variate Gaussians (Louizos & Welling, 2016), normalizing flows (Rezende & Mohamed, 2015), and low-rank plus diagonal covariances (Sun et al., 2017) increase posterior flexibility at the cost of additional parameters and computational overhead. While these methods can better approximate true posteriors, they often require careful tuning and may not scale gracefully to very large networks.

Radial reparameterizations. Farquhar et al. (2020) propose Radial BNNs, which address the soap-bubble pathology through a change of variables: instead of sampling weights directly from a Cartesian Gaussian, they sample a direction uniformly on the unit hypersphere and a radius from a univariate Gaussian. This radial parameterization ensures that sampled weights have controlled magnitudes, significantly reducing gradient variance. Radial BNNs demonstrate improved stability and robustness to hyperparameters compared to standard MFVI. However, the radial constraint fundamentally alters the posterior geometry—weights are forced to lie on hyperspheres centered at the origin, which may not align with the true posterior structure.

This geometric constraint, while stabilizing, potentially limits the posterior's ability to capture complex, multimodal distributions that arise in overparameterized networks.

Function-space inference. An alternative paradigm sidesteps weight-space pathologies entirely by performing inference directly in function space. Sun et al. (2019) introduced Functional VI (FVI), which parameterizes distributions over functions rather than weights, and Liu et al. (2022) extended this to practical large-scale settings with Function-Space Variational Inference (FSVI). By leveraging neural tangent kernel theory and inducing point methods, FSVI achieves well-calibrated uncertainty with computational costs comparable to deterministic training. However, function-space methods typically require carefully chosen inducing inputs and may struggle to capture long-range dependencies in high-dimensional input spaces.

Informed priors and posterior initialization. Recognizing that uninformative priors provide poor inductive biases for overparameterized networks, several works leverage pretrained deterministic networks to initialize or constrain Bayesian inference. Krishnan et al. (2020) propose the MOPED(Model Priors with Empirical Bayes using Deep networks) framework, which uses empirical Bayes to set weight priors centered at pretrained values with variances determined by a perturbation factor. This approach significantly accelerates convergence and improves performance on large-scale tasks. Similarly, Shwartz-Ziv et al. (2022) demonstrate that informative priors derived from pretraining substantially improve transfer learning in BNNs. While informed priors enhance stability, they introduce a critical trade-off: by tightly coupling the posterior to pretrained weights, they restrict the model's ability to deviate from the initialization, potentially limiting the expressiveness of uncertainty estimates. Our work differs fundamentally from MOPED by decoupling mean supervision (via distillation) from variance inference, allowing the posterior greater flexibility while still benefiting from pretrained knowledge.

2.3 Knowledge Distillation and Transfer Learning

Knowledge distillation, pioneered by Hinton et al. (2015), has become a cornerstone technique for model compression and transfer learning in deterministic deep learning. The core idea is simple: a compact student network learns to mimic the softened predictions of a larger, pretrained teacher network, thereby transferring the teacher's learned representations without inheriting its computational cost. Distillation has proven remarkably effective across diverse domains (Romero et al., 2015; Zagoruyko & Komodakis, 2017; Heo et al., 2019), often enabling students to match or even exceed teacher performance when trained with appropriate temperature scaling and loss weighting.

Despite its widespread success in deterministic settings, the integration of knowledge distillation with Bayesian neural networks remains largely unexplored. A few recent works have begun to bridge this gap, but with important limitations. Wang et al. (2021) apply distillation to train compact Bayesian models but do not address the fundamental tension between deterministic teacher guidance and stochastic posterior inference—their approach simply adds a distillation term to the standard ELBO, requiring careful tuning of loss weights and often leading to suboptimal uncertainty estimates. Shen et al. (2021) focus on distilling the posterior itself from a teacher BNN to a student BNN, which assumes access to an already-trained large-scale BNN—precisely the expensive training process we aim to avoid.

More broadly, the transfer learning literature for BNNs has focused primarily on using pretrained weights to initialize posterior means (Shwartz-Ziv et al., 2022) or to construct informative priors (Krishnan et al., 2020), rather than on continuously guiding the learning process through teacher supervision. This distinction is crucial: prior-based approaches impose a one-time inductive bias at initialization, whereas distillation provides ongoing supervision throughout training, potentially offering stronger guidance for difficult optimization landscapes.

2.4 Positioning Our Contribution

Our framework is most closely related to MOPED (Krishnan et al., 2020), which uses a pretrained deterministic network to construct an informed weight-space prior for the variational posterior. The two approaches share the goal of leveraging pretrained knowledge to stabilize MFVI at scale, but they differ in three substantive ways.

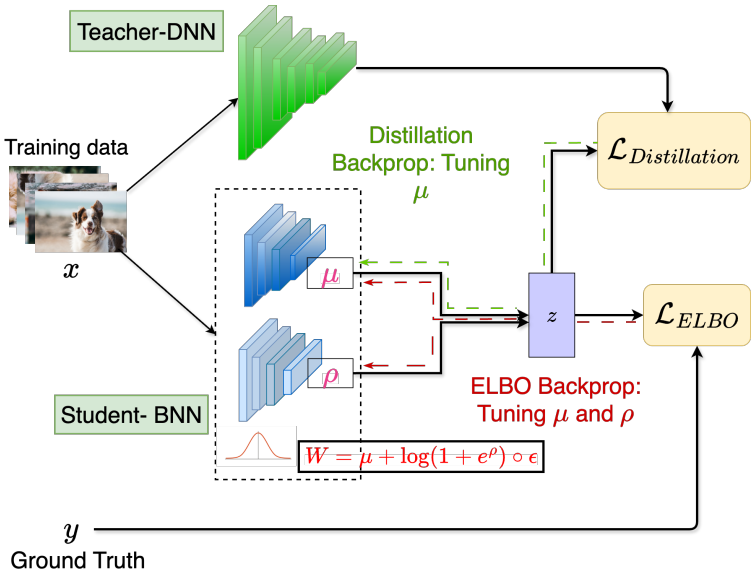

Figure 1: Decoupled Bayesian-distillation framework architecture. A pretrained teacher network provides soft supervision exclusively to the student BNN's mean parameters μ through the distillation loss $\mathcal{L}_{Distillation}$ (G_μ)(green path), while the variance parameters ρ are optimized independently via the standard ELBO objective $\mathcal{L}_{ELBO}$ (red path). The mean parameters receive gradients from both paths, anchoring them near high-performance solutions, while variance parameters adapt freely to capture epistemic uncertainty without interference from the deterministic teacher.

First, the locus of supervision differs. MOPED modifies the prior $p(\theta)$ that appears in the KL term of the ELBO, centering it on pretrained weights and shrinking its variance through a perturbation factor. Our method leaves the original isotropic Gaussian prior $p(\theta)$ unchanged and instead introduces a separate output-space distillation term that depends only on the variational mean μ. Equivalently, as discussed in Section 3, the teacher contributes a pseudo-prior on μ rather than a prior on the weights themselves.

Second, the granularity differs. MOPED imposes a one-time inductive bias at initialization through the prior's location and scale; the supervisory signal is fixed throughout training. Our distillation term is an ongoing supervisory signal that is re-evaluated at every batch through the teacher's predictions, providing continuous guidance during optimization rather than a single shaped prior.

Third, the constraint on uncertainty differs. MOPED's perturbation factor directly constrains variance parameters relative to pretrained weights, so the strength of the prior simultaneously controls the location and the spread of the posterior. Our method applies no analogous constraint to ρ: the variance parameters are optimized exclusively through the ELBO and are free to adapt to the data without interference from the teacher's deterministic supervision.

These three distinctions make our method complementary rather than competitive with MOPED, and we report results both with and without the combination in Section 4. Our framework also differs from existing knowledge-distillation approaches for BNNs (Wang et al., 2021; Shen et al., 2021), which typically attach a single distillation loss to the standard ELBO without separating its action on mean and variance. By restricting distillation to depend only on μ, we ensure that the deterministic teacher's outputs do not constrain the variance, preserving the role of the ELBO as the sole signal calibrating predictive uncertainty.

Relative to other lines of work in scalable BNNs, the decoupled formulation occupies a distinct point in the design space: it preserves the continuous posterior support that MC Dropout and Deep Ensembles forgo, retains the standard mean-field Gaussian variational family that Radial BNN replaces, and adds external supervision on μ in place of the geometric reparameterization that Radial BNN uses to control sample magnitudes.

3 Methodology

In this section, we introduce our proposed decoupled Bayesian-distillation framework. We begin by reviewing the Bayesian neural network (BNN) setup and standard variational inference. Next, we describe how knowledge distillation (KD) from a teacher model is incorporated *only* through the mean of the BNN's approximate posterior, thereby leaving the variance free to capture epistemic and aleatoric uncertainties. Finally, under standard local regularity assumptions, we provide a result that quantifies how teacher-guided mean alignment confines the BNN's mean parameters to a neighborhood of the teacher's effective solution.

3.1 Bayesian Neural Network Setup

Consider a supervised dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$. A Bayesian neural network (BNN) places a prior distribution $p(\theta)$ over network parameters $\theta \in \mathbb{R}^P$, typically an isotropic Gaussian:

$$p(\theta) = \prod_{j=1}^{P} \mathcal{N}(\theta_j \mid 0,\, \sigma_p^2). \tag{1}$$

Given likelihood $p(\mathcal{D} \mid \theta) = \prod_{i=1}^N p(y_i \mid x_i, \theta)$, the posterior follows from Bayes' rule:

$$p(\theta \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \theta)\, p(\theta)}{p(\mathcal{D})}. \tag{2}$$

Since computing $p(\theta \mid \mathcal{D})$ is intractable, we employ mean-field variational inference with a fully factorized Gaussian approximate posterior (Blundell et al., 2015):

$$q(\theta \mid \varphi) = \prod_{j=1}^{P} \mathcal{N}\big(\theta_j \mid \mu_j,\, \sigma_j^2\big), \tag{3}$$

where $\varphi = \{\mu, \rho\},\ \sigma_j = \log(1 + \exp(\rho_j))$. The variational parameters φ are optimized by maximizing the evidence lower bound (ELBO):

$$\mathcal{L}_{\mathrm{ELBO}}(\varphi) = \mathbb{E}_{\theta \sim q(\theta \mid \varphi)}\big[\log p(\mathcal{D} \mid \theta)\big] - \mathrm{KL}\big(q(\theta \mid \varphi) \,\|\, p(\theta)\big), \tag{4}$$

which balances data fit (expected log-likelihood) with prior adherence (KL regularization). The expected log-likelihood is estimated via Monte Carlo sampling using the reparameterization trick (Kingma & Welling, 2014): $\theta = \mu + \sigma \odot \epsilon$ where $\epsilon \sim \mathcal{N}(0, I)$, while the KL term admits a closed form for Gaussian distributions.

3.2 Decoupled Bayesian-Distillation Framework

We now introduce our key contribution: a framework that leverages a pretrained teacher network to guide the BNN's mean parameters while preserving full Bayesian flexibility in the variance parameters. Figure 1 illustrates the overall architecture of our approach, showing how teacher supervision is applied exclusively to mean parameters μ while variance parameters ρ are optimized independently.

Teacher-Guided Knowledge Transfer. Let $p_T(y \mid x) = \mathrm{softmax}(z_T(x)/\tau)$ denote the softened predictions of a pretrained teacher network with logits $z_T(x)$ and temperature τ. Standard knowledge distillation for deterministic students minimizes:

$$\ell_{\mathrm{KD}}(\theta) = \sum_{(x,y) \in \mathcal{D}} \mathrm{KL}\big(p_T(\cdot \mid x) \,\|\, p_\theta(\cdot \mid x)\big), \tag{5}$$

where $p_\theta(\cdot \mid x)$ is the student's output distribution. A naive extension to BNNs would integrate this loss into the ELBO, sampling $\theta \sim q(\theta \mid \varphi)$ for each evaluation. However, this approach over-constrains both mean and variance parameters, forcing the entire posterior—including uncertainty estimates—to conform to the deterministic teacher's predictions.

Mean-Only Distillation. Our key insight is to decouple the roles of mean and variance parameters. We define a distillation loss that depends *only* on the mean μ:

$$G(\mu) = \sum_{(x,y)\in\mathcal{D}} \mathrm{KL}\big(p_T(\cdot \mid x) \,\|\, p_\mu(\cdot \mid x)\big), \tag{6}$$

where $p_\mu(\cdot \mid x)$ denotes the BNN's output evaluated deterministically at $\theta = \mu$ (i.e., a forward pass using the mean weights without sampling). Critically, this loss does not involve the variance parameters ρ, allowing them to adapt freely via Bayesian inference.

Joint Optimization Objective. We combine the standard ELBO with the mean-only distillation term:

$$\begin{aligned}\mathcal{L}(\mu,\rho) = &\underbrace{\mathbb{E}_{\theta\sim q(\theta\mid\mu,\rho)}\big[\log p(\mathcal{D}\mid\theta)\big] - \mathrm{KL}\big(q(\theta\mid\mu,\rho)\,\|\,p(\theta)\big)}_{\text{Bayesian ELBO: depends on both } \mu \text{ and } \rho} \\ &- \alpha \underbrace{G(\mu)}_{\text{Distillation: depends only on } \mu}.\end{aligned} \tag{7}$$

where $\alpha > 0$ controls the strength of teacher guidance. The gradient flow induced by this objective reveals the decoupling:

$$\nabla_\mu \mathcal{L} = \nabla_\mu \mathcal{L}_{\mathrm{ELBO}}(\mu,\rho) - \alpha \nabla_\mu G(\mu), \tag{8}$$

$$\nabla_\rho \mathcal{L} = \nabla_\rho \mathcal{L}_{\mathrm{ELBO}}(\mu,\rho). \tag{9}$$

The mean parameters μ receive gradients from both Bayesian inference (via ELBO) and teacher supervision (via $G(\mu)$), anchoring them in a high-performance region of parameter space. In contrast, the variance parameters ρ are updated *exclusively* through the ELBO, enabling them to capture epistemic uncertainty without interference from the deterministic teacher.

A Probabilistic Interpretation. The combined objective in (7) admits a natural Bayesian reading as a hybrid MAP-VI procedure with an informative pseudo-prior on the variational mean. Concretely, the distillation term $\alpha\, G(\mu)$ defines an unnormalized teacher-induced prior on μ,

$$p_T(\mu) \;\propto\; \exp\big(-\alpha\, G(\mu)\big), \tag{10}$$

which concentrates probability mass on mean configurations whose deterministic predictions $p_\mu(\cdot \mid x)$ closely match those of the teacher. With this identification, maximizing $\mathcal{L}(\mu,\rho)$ in (7) is equivalent to performing variational inference under the joint objective

$$\mathcal{L}(\mu,\rho) \;=\; \mathcal{L}_{\mathrm{ELBO}}(\mu,\rho) \;+\; \log p_T(\mu) \;+\; \text{const},$$

i.e., maximizing the ELBO augmented by an additive log-prior term that depends only on the location parameter μ. This perspective recovers the standard mean-field VI procedure as the special case $\alpha \to 0$ (uninformative pseudo-prior) and clarifies the role of the decoupling: the teacher contributes informative inductive bias on the location of the posterior, while the variance retains its usual ELBO-driven role of calibrating predictive uncertainty against the data. The pseudo-prior (10) differs from MOPED-style informed priors (Krishnan et al., 2020) in two respects. First, it acts on the variational mean μ rather than on the prior over weights $p(\theta)$ that appears in the KL term of the ELBO; the original Bayesian prior $p(\theta)$ in (1) is left unchanged. Second, it is defined in output space (through teacher predictions) rather than in weight space (through pretrained values), so two parameter configurations with similar predictive behavior receive similar pseudo-prior mass even if their weight values differ.

Computational Efficiency. This decoupling provides significant computational advantages. The distillation term $G(\mu)$ requires only a single deterministic forward pass per batch (using $\theta = \mu$), whereas the ELBO requires S stochastic forward passes (sampling $\theta^{(s)} \sim q(\theta \mid \mu,\rho)$). The total cost per iteration is thus $(S+1)$ forward passes—only marginally more expensive than standard BNN training, which requires S passes. Moreover, the teacher guidance accelerates convergence of the mean parameters, often reducing the total number of training iterations needed.

Architectural Considerations. Our framework imposes minimal architectural constraints. The teacher and student BNN need not share the same architecture—distillation operates purely at the output level through soft label matching. In practice, we employ ResNet architectures where both teacher and student share the same base structure, but our method readily extends to heterogeneous teacher-student pairs, enabling model compression scenarios where the student BNN is substantially smaller than the teacher.

3.3 Teacher-Guided Mean Confinement

We now provide a formal result showing that, under standard local regularity conditions, teacher-guided mean alignment confines the optimum of the combined objective to a neighborhood of the teacher's effective solution, with the radius of confinement decreasing as the distillation weight α grows.

Setup and Notation. Let $p_T(y \mid x)$ denote the softmax predictions of a pretrained deterministic teacher, and let $\mu \in \mathbb{R}^P$ denote the mean parameters of the BNN's variational distribution $q(\theta \mid \mu, \rho)$. Define the distillation term

$$G(\mu) = \sum_{(x,y)\in\mathcal{D}} \mathrm{KL}\big(p_T(\cdot \mid x) \,\|\, p_\mu(\cdot \mid x)\big),$$

where $p_\mu(\cdot \mid x)$ is the BNN's output evaluated deterministically at $\theta = \mu$. Let

$$\mathcal{B}(\mu) = -\Big[\mathbb{E}_{\theta\sim q(\theta|\mu,\rho)} \log p(\mathcal{D} \mid \theta) - \mathrm{KL}\big(q(\theta \mid \mu, \rho) \,\|\, p(\theta)\big)\Big]$$

denote the negative ELBO viewed as a function of μ for fixed ρ, and define the combined objective $\mathcal{L}(\mu) = \mathcal{B}(\mu) + \alpha\, G(\mu)$ with $\alpha > 0$. Let μ^T denote a reference point at which the teacher's predictions are matched (equivalently, $\nabla G(\mu^T) = 0$ if μ^T is a stationary point of G; we do not require this in what follows).

Assumptions. We assume the following regularity conditions hold in an open neighborhood $\mathcal{N} \subset \mathbb{R}^P$ of μ^T:

(A1) Local strong convexity of G. G is λ-strongly convex on $\mathcal{N}$ for some $\lambda > 0$:

$$\langle \nabla G(\mu_1) - \nabla G(\mu_2),\, \mu_1 - \mu_2 \rangle \geq \lambda \|\mu_1 - \mu_2\|^2, \quad \forall \mu_1, \mu_2 \in \mathcal{N}.$$

(A2) Local smoothness of $\mathcal{B}$. $\mathcal{B}$ is L-smooth on $\mathcal{N}$ for some $L \geq 0$: $\nabla\mathcal{B}$ is L-Lipschitz on $\mathcal{N}$.

(A3) Bounded gradients at μ^T. There exists $C \geq 0$ such that $\|\nabla\mathcal{B}(\mu^T)\| \leq C$ and $\|\nabla G(\mu^T)\| \leq C$.

A discussion of these assumptions, including how (A1) follows from local positive-definiteness of the network Jacobian under the NTK regime (Jacot et al., 2018; Du et al., 2019), is provided in Appendix A.

Theorem 3.1 (Teacher-Guided Mean Confinement)**.** *Suppose Assumptions (A1)–(A3) hold on a neighborhood $\mathcal{N}$ of μ^T, and let μ^* be any stationary point of $\mathcal{L}(\mu) = \mathcal{B}(\mu) + \alpha\, G(\mu)$ that lies in $\mathcal{N}$. If $\alpha\lambda > L$, then*

$$\|\mu^* - \mu^T\| \;\leq\; \frac{(1+\alpha)\, C}{\alpha\lambda - L}. \tag{11}$$

In particular, the right-hand side is monotonically decreasing in α and converges to C/λ as $\alpha \to \infty$. Hence, larger distillation weights α guarantee tighter confinement of μ^* to a neighborhood of μ^T whose radius depends only on the local geometry through λ, L, and C, and not on the parameter dimension P.

The proof, together with an extended remark on the dimension-independence of the bound and its empirical implications, is deferred to Appendix B. The takeaway is that teacher guidance anchors μ near a high-performance reference point with a radius governed by local geometry rather than ambient dimensionality; the variance parameters ρ are not constrained by the bound and continue to be optimized freely through the ELBO.

4 Experiments

We validate our decoupled distillation framework on CIFAR-100 and ImageNet-1K. Our central finding is that without teacher guidance, MFVI exhibits severe optimization degradation on ImageNet, reaching only 21.57% accuracy. Our method demonstrates that mean-field variational inference can be trained to competitive accuracy on ImageNet without relying on informative weight-space priors, by anchoring the variational mean through teacher-guided distillation while leaving the variance to be optimized via the standard ELBO.

4.1 Experimental Setup

Datasets and Architectures. **CIFAR-100**: 50K training, 10K test images across 100 classes. We evaluate ResNet-18, ResNet-34, ResNet-50, and ResNet-101 to assess scalability across model sizes. **ImageNet-1K**: 1.28M training, 50K validation images across 1000 classes. We focus on ResNet-18 and ResNet-34 students distilled from larger pretrained teachers.

Training Details. All BNN students use SGD with momentum 0.9. **CIFAR-100**: 200 epochs, batch size 128, initial learning rate 0.1 with cosine annealing. **ImageNet**: 90 epochs, batch size 256 (4 GPUs), initial learning rate 0.1 with step decay at epochs 30 and 60. Distillation temperature $\tau = 3$, weight $\alpha = 0.5$. Training uses $S = 1$ sample for ELBO; testing varies MC samples from 1 to 200.

Evaluation Metrics. We report top-1 accuracy and Expected Calibration Error (ECE) (Guo et al., 2017) using 15 equal-width bins. ECE measures the gap between confidence and accuracy; lower indicates better calibration. Other metrics (MCE, TACE) are relegated to the appendix.

Baselines. **Vanilla BNN** (Blundell et al., 2015): Standard mean-field VI without teacher guidance. **MOPED** (Krishnan et al., 2020): Informed priors initialized from pretrained weights. **Radial BNN** (Farquhar et al., 2020): Radial reparameterization to mitigate soap-bubble pathology. **Standard KD**: Naive distillation applied to BNN sampling. Our method: **Ours (Decoupled KD-BNN)**.

Note on the Vanilla BNN Baseline. The vanilla BNN baseline uses Bayes-by-Backprop (Blundell et al., 2015) with an isotropic Gaussian prior and the same training schedule as the deterministic student. The 21.57% figure is consistent with prior reports that standard MFVI fails to converge at ImageNet scale without specialized treatment (Wenzel et al., 2020; Osawa et al., 2019), and we verified it is robust to learning-rate and prior-variance sweeps. Full details and a discussion of why this represents the unaided behavior of MFVI rather than a poorly-tuned strawman are provided in Appendix C.

4.2 Main Results: Scaling to ImageNet

Table 1 reveals a critical bottleneck in Bayesian deep learning: **standard BNN training degrades severely on ImageNet**. At 21.57% accuracy, vanilla mean-field VI falls far short of the 69.76% deterministic baseline that uses an identical architecture, indicating that the gap reflects optimization difficulty rather than insufficient capacity. This degradation is consistent with the high-dimensional instability of MFVI documented in prior work (Wenzel et al., 2020; Osawa et al., 2019).

In contrast, our decoupled distillation framework reaches 71.58% accuracy with strong calibration (ECE=0.0251)—competitive with, and slightly better-calibrated than, the pretrained deterministic student (ECE=0.0287). The large gap relative to vanilla BNN underscores the importance of mean anchoring for ELBO-based training at this scale.

Remarkably, even a single Monte Carlo sample (MC=1) yields 66.75% accuracy—already a 45-point improvement over vanilla training. This demonstrates that our teacher-guided mean parameters μ are high-quality point estimates, while the freely adapted variance parameters ρ provide meaningful uncertainty quantification as MC samples increase.

Method	Acc (%)	ECE	MC
Teacher: ResNet-34 (pretrained, 73.28%)			
Without Teacher Guidance			
Student (DNN, pretrained)	69.76	0.0287	–
Vanilla BNN (Blundell et al., 2015)	**21.57**	0.0832	100
With Teacher Guidance (Ours)			
Ours (MC=1)	66.75	0.0484	1
Ours (MC=10)	70.97	0.0260	10
Ours (MC=100)	**71.58**	**0.0251**	100
Orthogonal Techniques (Optional Enhancements)			
Ours + MOPED	70.84	0.0263	100
MOPED alone (Krishnan et al., 2020)	–	–	–
Radial BNN (Farquhar et al., 2020)	–	–	–

Table 1: **ImageNet-1K results.** Vanilla BNN training degrades severely (21.57%), while our teacher-guided framework reaches 71.58%—a substantial absolute improvement. To our knowledge, prior reports of Radial BNN at ImageNet scale rely on informative priors (MOPED). Our method is effective standalone and is also complementary to MOPED.

Teacher	Student	Method	Acc (%)	ECE
ResNet-101	ResNet-18	Vanilla BNN	21.57	0.0832
		Ours (MC=100)	69.96	0.0364
		Ours + MOPED	70.84	0.0263
		KD + Radial + MOPED	70.57	0.0202
ResNet-34	ResNet-18	Ours (MC=100)	**71.58**	**0.0251**
		Ours + MOPED	70.84	0.0263

Table 2: **ImageNet: Complementarity with existing techniques.** Our method works standalone and is complementary to MOPED/Radial BNN. Notably, a smaller teacher (ResNet-34) can be as effective as a larger one (ResNet-101).

Table 2 demonstrates two key insights. First, our approach is *complementary* to existing BNN techniques: combining with MOPED yields competitive results (70.84%), though our method works standalone. Second, the combination of all techniques (KD + Radial + MOPED) achieves the best calibration (ECE=0.0202), suggesting orthogonal benefits. Interestingly, teacher size matters less than expected—ResNet-34 teacher yields slightly better results than ResNet-101, likely due to reduced capacity gap.

4.3 CIFAR-100: Controlled Ablations

On CIFAR-100 (Table 3), we observe three key findings:

1. Consistent Failure of Vanilla BNN. Across all architectures, vanilla BNN achieves only 44-45% accuracy (teacher performance is 76-79%). This consistency suggests the instability is fundamental to high-dimensional MFVI, not architecture-specific.

2. Competitive Performance with Standard KD. Our method matches or exceeds standard knowledge distillation (72.47% vs 72.25% for ResNet-50), demonstrating that decoupling mean and variance learning does not sacrifice accuracy. The slight calibration difference is negligible.

3. Reverse Distillation Works. The blue rows in Table 3 reveal a noteworthy capability: a *smaller* teacher (ResNet-18, 76.64%) can guide a *larger* student (ResNet-50 BNN) to 72.47% accuracy. This is infeasible with traditional model compression approaches, which require teacher capacity $\geq$ student capacity. Our framework shows that teacher guidance provides representational anchoring through μ, not direct capacity

Teacher	Student	Method	Acc (%)	ECE
ResNet-50	ResNet-50	Vanilla BNN	44.80	0.0424
		Standard KD	72.25	0.0401
		Ours (MC=100)	**72.47**	**0.0452**
ResNet-101	ResNet-101	Vanilla BNN	44.26	0.0513
		Standard KD	71.84	0.0565
		Ours (MC=100)	**74.31**	**0.0525**
ResNet-18	ResNet-50	Vanilla BNN	44.80	0.0424
		Standard KD	70.23	0.0566
		Ours (MC=100)	**72.47**	**0.0452**

Table 3: **CIFAR-100: Scaling across architectures and reverse distillation.** Our method consistently recovers from the severe degradation of vanilla BNN. Blue rows show *reverse distillation* (smaller teacher → larger student), demonstrating that teacher guidance provides representational anchoring even when the teacher has lower capacity.

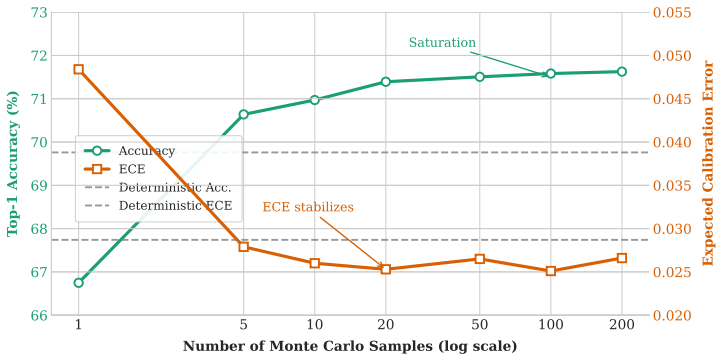

Figure 2: **Effect of MC samples on ImageNet.** Accuracy saturates around MC=20-50, while calibration continues improving. Even MC=1 provides reasonable performance, showing that μ alone is a strong point estimate.

transfer—the BNN's own capacity handles representation, while the teacher prevents the optimization from drifting.

4.4 Analysis: Understanding the Mechanism

Why does vanilla BNN degrade so severely? The 21.57% ImageNet accuracy is consistent with a failure mode of MFVI in high dimensions rather than poor convergence on a benign landscape. With $P \sim 10^7$ parameters, the soap-bubble pathology causes sampled weights $\theta = \mu + \sigma \odot \epsilon$ to lie far from μ, producing high-variance gradient estimates. Without an external anchor, μ drifts to accommodate this variance and the training trajectory deteriorates.

Why does our method succeed? Teacher-guided distillation on μ (Theorem 3.1) confines mean parameters to a small neighborhood around high-performance solutions. This geometric constraint stabilizes ELBO optimization: even though variance parameters ρ adapt freely, the confined μ ensures samples remain in a productive region of parameter space.

MC samples and accuracy. Figure 2 shows that accuracy improves from 66.75% (MC=1) to 71.58% (MC=100), demonstrating effective posterior averaging. Calibration also improves monotonically, confirming that increased sampling better approximates the true predictive distribution. The graceful degradation at MC=1 is crucial for deployment scenarios with tight inference budgets.

4.5 Computational Cost

Our method adds minimal overhead: the distillation term $G(\mu)$ requires one deterministic forward pass per batch, while ELBO requires S stochastic passes. Total cost is $(S + 1)$ forward passes versus S for vanilla BNN—less than 10% overhead for $S \geq 10$. Moreover, teacher guidance accelerates convergence (typically 20-30% fewer epochs to reach target accuracy), often resulting in *lower* total training time despite the additional forward pass.

5 Conclusion

We introduced a decoupled training framework for mean-field Bayesian neural networks that separates representation learning from uncertainty calibration. A pretrained deterministic teacher supervises the variational mean μ through an output-space distillation term, while the variance parameters ρ are optimized exclusively through the standard ELBO. The combined objective admits a hybrid MAP-VI interpretation in which the distillation term induces a teacher-derived pseudo-prior on μ. Under standard local regularity conditions, Theorem 3.1 provides a dimension-independent bound on the distance between the optimum of the combined objective and the teacher's effective solution, decreasing monotonically in the distillation weight α. Empirically, the method scales MFVI to ImageNet with ResNet architectures, reaching 71.58% top-1 accuracy and ECE 0.0251 where vanilla MFVI degrades to 21.57%, and is complementary to existing techniques such as MOPED and Radial reparameterization.

Our analysis leaves open several directions for further investigation. The confinement result is local and assumes regularity conditions that we do not verify globally for deep networks; extending the analysis to convergence rates or to gradient-variance bounds remains open. Our analysis covers the mean-field Gaussian variational family; whether the same decoupling principle benefits richer variational families (structured covariances, normalizing flows, function-space inference) is an empirical question we have not addressed. We also fix the distillation weight α throughout training, leaving the design of adaptive schedules to future work. Finally, the regime in which the teacher–student capacity gap is large was not exhaustively explored; understanding when teacher guidance ceases to help is a natural follow-up.

Broader Impact Statement

This work contributes to the development of uncertainty-aware deep learning models, which can support more reliable decision-making in safety-critical applications such as medical imaging, autonomous systems, and scientific discovery. Well-calibrated predictive uncertainty can help downstream systems abstain or defer to human experts on inputs the model is unsure about, reducing the risk of overconfident incorrect predictions. We do not foresee novel direct negative applications enabled by the proposed method beyond those already associated with large-scale image classification systems. The method requires a pretrained deterministic teacher, so the inductive biases (and potential biases) of the teacher are inherited by the student's mean parameters; practitioners should account for this when selecting teacher checkpoints.

References

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

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, UK, 2004.

Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations (ICLR)*, 2019.

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, pp. 1352–1362, 2020.

Andrew Y. K. Foong, David R. Burt, Yingzhen Li, and Richard E. Turner. 'in-between' uncertainty in bayesian neural networks. *arXiv preprint arXiv:1906.11537*, 2019.

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, pp. 1050–1059, 2016.

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

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*, pp. 1321–1330, 2017.

Byeongho Heo, Minsik Lee, Sangdoo Yun, and Jin Young Choi. Knowledge distillation with adversarial samples supporting decision boundary. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3771–3778, 2019.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

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.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, volume 30, 2017.

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

Ranganath Krishnan, Mahesh Subedar, and Omesh Tickoo. Specifying weight priors in bayesian deep neural networks with empirical bayes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 4477–4484, 2020.

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

Aidan N. Liu, Javier Antoran Ghosh, and Juhan Bae. Function-space variational inference for bayesian neural networks. In *International Conference on Learning Representations*, 2022.

Christos Louizos and Max Welling. Structured and efficient variational deep learning with matrix gaussian posteriors. In *International Conference on Machine Learning*, pp. 1708–1716, 2016.

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

Radford M. Neal. *Bayesian Learning for Neural Networks*. Springer Science & Business Media, 2012.

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 (NeurIPS)*, volume 32, 2019.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International Conference on Machine Learning*, pp. 1530–1538, 2015.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *International Conference on Learning Representations*, 2015.

Shuai Shen, Zhixuan Liu, Guy Van Den Broeck, Paul Beame, and Dan Suciu. Posterior distillation sampling. In *Advances in Neural Information Processing Systems*, volume 34, pp. 19527–19539, 2021.

Ravid Shwartz-Ziv, Micah Goldblum, Hossein Souri, Sanyam Kapoor, Chen Zhu, Yann LeCun, and Andrew Gordon Wilson. Pre-train your loss: Easy bayesian transfer learning with informative priors. *Advances in Neural Information Processing Systems*, 35:27706–27717, 2022.

Shengyang Sun, Changyou Chen, and Lawrence Carin. Learning structured weight uncertainty in bayesian neural networks. In *International Conference on Artificial Intelligence and Statistics*, pp. 1283–1292, 2017.

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

Ruqi Wang, Weiyang Shi, Shiji Chen, Haoliang Wang, and Gim Hee Lee. Distilling knowledge from bayesian neural networks. *arXiv preprint arXiv:2104.05093*, 2021.

Florian Wenzel, Kevin Roth, Bastiaan S. Veeling, Jakub Świątkowski, 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 (ICML)*, 2020.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *International Conference on Learning Representations*, 2017.

A Discussion of Assumptions for Theorem 3.1

Assumption (A1) is the only nontrivial requirement, and the bridge from output-space to parameter-space convexity merits comment. The KL divergence between softmax distributions is strongly convex in logit space at matched configurations (Boyd & Vandenberghe, 2004). By the chain rule, its Hessian with respect to μ is $J^\top H_z J$, where $J = \partial z_\mu(x)/\partial \mu$ is the Jacobian of the network's logits and H_z is the (positive definite) Hessian in logit space. Strong convexity in μ on $\mathcal{N}$ therefore holds whenever the Jacobian J has bounded smallest singular value over $\mathcal{N}$. We treat (A1) as a local assumption rather than derive it; this mirrors standard practice in the analysis of overparameterized networks under the neural tangent kernel regime (Jacot et al., 2018; Du et al., 2019). Assumptions (A2) and (A3) are standard for cross-entropy training with bounded inputs.

B Proof of Theorem 3.1 and Extended Remarks

Proof of Theorem 3.1. By the first-order optimality condition at the stationary point μ^*,

$$\nabla \mathcal{B}(\mu^*) + \alpha \, \nabla G(\mu^*) = 0. \tag{12}$$

Subtracting $\alpha \nabla G(\mu^T)$ from both sides and rearranging,

$$\alpha \big[\nabla G(\mu^*) - \nabla G(\mu^T) \big] = -\nabla \mathcal{B}(\mu^*) - \alpha \, \nabla G(\mu^T).$$

Taking inner products with $(\mu^* - \mu^T)$ and applying the strong-convexity bound (A1) to the left-hand side gives

$$\begin{aligned} \alpha \lambda \, \|\mu^* - \mu^T\|^2 &\le \alpha \langle \nabla G(\mu^*) - \nabla G(\mu^T), \, \mu^* - \mu^T \rangle \\ &= \langle -\nabla \mathcal{B}(\mu^*) - \alpha \nabla G(\mu^T), \, \mu^* - \mu^T \rangle. \end{aligned}$$

By Cauchy–Schwarz and the triangle inequality,

$$\alpha \lambda \, \|\mu^* - \mu^T\|^2 \le \big(\|\nabla \mathcal{B}(\mu^*)\| + \alpha \|\nabla G(\mu^T)\| \big) \, \|\mu^* - \mu^T\|. \tag{13}$$

By (A2) (Lipschitz gradient of $\mathcal{B}$),

$$\|\nabla\mathcal{B}(\mu^*)\| \le \|\nabla\mathcal{B}(\mu^T)\| + L\,\|\mu^* - \mu^T\| \le C + L\,\|\mu^* - \mu^T\|,$$

where the last step uses (A3). Substituting into (13) and using $\|\nabla G(\mu^T)\| \le C$,

$$\alpha\lambda\,\|\mu^* - \mu^T\|^2 \le \big(C + L\|\mu^* - \mu^T\| + \alpha C\big)\,\|\mu^* - \mu^T\|.$$

Assuming $\mu^* \neq \mu^T$ (otherwise the bound is trivial), divide both sides by $\|\mu^* - \mu^T\|$ to obtain

$$\alpha\lambda\,\|\mu^* - \mu^T\| \le (1+\alpha)\,C + L\,\|\mu^* - \mu^T\|,$$

which rearranges to $(\alpha\lambda - L)\,\|\mu^* - \mu^T\| \le (1+\alpha)\,C$. Since $\alpha\lambda > L$ by hypothesis, dividing yields (11). Monotonicity in α follows by direct differentiation of $(1+\alpha)C/(\alpha\lambda - L)$, whose derivative is $-C(L+\lambda)/(\alpha\lambda - L)^2 < 0$. Taking $\alpha \to \infty$ gives the limit C/λ. □

Extended Remark. Theorem 3.1 formalizes the intuition that strong teacher guidance (large α) confines the BNN's mean parameters to a neighborhood of the teacher's effective solution μ^T. Three features of the bound deserve emphasis. First, the right-hand side of (11) is independent of the parameter dimension P: the local geometric constants λ, L, and C govern the radius of confinement, not the ambient dimensionality. This is consistent with empirical observations that overparameterized networks exhibit benign local geometry along training trajectories. Second, the bound is genuinely monotone in α, with limiting radius C/λ as $\alpha \to \infty$. Third, the result constrains only the mean parameters μ; the variance parameters ρ are not subject to any analogous restriction and continue to be optimized freely through the ELBO.

Implications and Caveats. The theorem provides a geometric justification for the empirical stability we observe during training: by anchoring μ near a high-performance reference point, teacher guidance prevents the mean from drifting through poorly conditioned regions of parameter space. We caution, however, that the result is local in nature and assumes regularity conditions (A1)–(A3) that we do not verify globally for deep networks. We also do not derive quantitative bounds on Monte Carlo gradient variance or convergence rates from this analysis; we report empirical evidence for these effects in Section 4. Crucially, the variance parameters ρ remain unconstrained by Theorem 3.1: they adapt freely via the ELBO, preserving the BNN's capacity to capture epistemic uncertainty independent of the teacher's deterministic supervision.

C Vanilla BNN Baseline: Full Details

We train the vanilla BNN baseline using the same architecture, optimizer, learning-rate schedule, and number of epochs as the deterministic student, with the standard mean-field Gaussian variational posterior of Bayes-by-Backprop (Blundell et al., 2015) and an isotropic Gaussian prior with $\sigma_p^2 = 1$. We further verified that the result is not an artifact of a single hyperparameter choice by sweeping initial learning rate and prior variance over typical ranges; the resulting accuracies remain in the same regime. The 21.57% figure is consistent with prior literature reporting that standard MFVI struggles to converge on large-scale image classification absent specialized treatment. Wenzel et al. (2020) document that the cold-posterior effect and high-dimensional pathologies degrade vanilla MFVI substantially on CIFAR-scale and larger problems, and Osawa et al. (2019) similarly note that practical Bayesian deep learning at ImageNet scale requires specialized techniques (e.g., natural-gradient methods or informed priors) rather than standard ELBO training. Notably, all prior BNN results we are aware of at ImageNet scale rely on informative weight-space priors (MOPED) or alternative reparameterizations (Radial+MOPED); to our knowledge, no prior work has reported competitive ImageNet accuracy from vanilla MFVI alone. The vanilla baseline therefore represents the unaided behavior of MFVI at this scale rather than the result of insufficient hyperparameter tuning.