BULLETIN OF COMPUTER AND DATA SCIENCES

The Cold Posterior Effect in Random Features Models: A Theoretical Explanation

Jun Li, Xiao Bai and Jin Zheng

The school of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing, China

Abstract

The "cold posterior effect"—where raising the Bayesian posterior to a power greater than 1 improves predictive performance—remains one of the most puzzling empirical phenomena in Bayesian deep learning. While numerous heuristic explanations have been proposed, a rigorous theoretical understanding remains elusive. In this paper, we provide the first theoretical analysis of this effect through the lens of random features regression. We prove that in the overparameterized regime, the posterior predictive distribution becomes systematically over-dispersed relative to the true risk of the maximum a posteriori (MAP) estimator. This miscalibration naturally suggests tempering the posterior to achieve better uncertainty quantification. Using recent asymptotic results for Bayesian random features models, we derive explicit conditions under which cold tempering improves frequentist coverage of credible sets and characterize the optimal temperature parameter. Our theoretical results are validated by numerical experiments and provide a mathematically grounded explanation for why cold posteriors work in practice.

Keywords: cold posterior effect, Bayesian deep learning, random features regression, posterior tempering, uncertainty calibration

1. Introduction

Bayesian methods offer a principled framework for uncertainty quantification in deep learning, providing a coherent way to combine prior knowledge with data and to obtain full predictive distributions rather than point estimates. In principle, when the model is correctly specified and the prior reasonably reflects our beliefs, the Bayesian posterior should be optimal in a precise decision-theoretic sense: it minimises posterior expected loss and yields well-calibrated predictive uncertainties. However, the practical deployment of Bayesian neural networks has revealed several puzzling empirical behaviors that challenge this idealized picture. Perhaps the most striking among them is the so-called "cold posterior effect" [1], where scaling the likelihood term in the Bayesian posterior by a factor $1/T$ with $T < 1$ (equivalently, raising the likelihood to a power $1/T$ and thus "cooling" the posterior) consistently improves predictive accuracy and calibration. This empirical preference for a *tempered* posterior, rather than the nominal Bayesian one with $T = 1$, has been observed across a variety of architectures, datasets, and approximate inference schemes [1, 2], and stands in direct tension with the theoretical optimality of the untempered posterior under the standard Bayesian assumptions.

A growing body of work has attempted to explain this phenomenon. One line of argument attributes the cold posterior effect to model misspecification [3]: if the data-generating process lies

Corresponding author: Jun Li (junli1987@proton.me)

outside the assumed model class, the true posterior need not be optimal for prediction, and tempering can act as a heuristic robustness correction. Another perspective emphasizes data curation and the interaction between training pipelines and the downstream Bayesian update [4], suggesting that heavy preprocessing or selection biases in curated datasets may effectively distort the likelihood, again making $T < 1$ appear favorable. A third explanation focuses on prior misspecification [2], pointing out that common architectural or weight priors in Bayesian deep learning (e.g., independent Gaussian priors on weights) may be poorly aligned with the inductive biases of trained networks, so that a colder posterior compensates for an overly diffuse or otherwise inappropriate prior. While these explanations are intuitively appealing and empirically supported in specific settings [1–4], they remain largely heuristic and are not grounded in a unified, rigorous asymptotic theory.

In parallel, the last few years have witnessed substantial progress in the theoretical understanding of overparameterized models, particularly in regression and classification with highly expressive function classes. Precise asymptotic characterizations of risk and generalization error have been obtained in a variety of high-dimensional limits, revealing phenomena such as double descent in test error as a function of model complexity [5]. These developments show that classical statistical intuition can fail dramatically in overparameterized regimes, and that careful asymptotic analysis is necessary to understand the behavior of modern models. Yet, despite this progress, the connection between overparameterization theory and the cold posterior effect has not been systematically explored: we lack a setting in which the Bayesian posterior, its tempered variants, and the corresponding frequentist risks can all be characterized in closed form and compared on equal footing.

In this work, we bridge this gap by analyzing the cold posterior effect within the mathematically tractable random features (RF) model for regression. Random features provide a simplified but rich proxy for wide neural networks, capturing key aspects of overparameterization while remaining amenable to precise asymptotic analysis. Building on the recent work of [6], which characterizes the asymptotics of Bayesian uncertainty estimation in RF regression, we show that in overparameterized regimes the standard Bayesian posterior systematically *overestimates* predictive uncertainty relative to the actual prediction error of the maximum a posteriori (MAP) estimator. In other words, the posterior predictive distribution is over-dispersed: its variance is larger than what is warranted by the frequentist risk of the underlying estimator. This misalignment between Bayesian and frequentist uncertainty measures provides a natural and conceptually simple justification for tempering the posterior: by effectively rescaling the likelihood, one can correct the overdispersion of the posterior predictive distribution and obtain uncertainty estimates that more closely track the true prediction error [1, 6].

1.1. Our Contributions

Against this backdrop, our contributions are as follows. Working in the high-dimensional random features regression model, we first provide a rigorous theoretical analysis showing that, in overparameterized regimes, the expected posterior predictive variance (PPV) of the standard Bayesian posterior exceeds the frequentist risk of the MAP predictor; this establishes that the nominal posterior is systematically over-dispersed and thus underconfident about its own predictions, even though it may perform well in terms of pointwise accuracy [6]. We then derive an explicit asymptotic characterization of an optimal temperature parameter $T^* < 1$ that aligns the posterior predictive variance with the MAP risk, thereby minimizing the discrepancy between Bayesian and frequentist uncertainty measures and offering a principled target for posterior tempering, in line with the empirical

observations of cold posteriors in deep models [1, 2]. Our analysis further reveals a phase transition in the behavior of this optimal temperature: depending on the signal-to-noise ratio (SNR) and the overparameterization ratio (features-to-samples), the value of T^* and its qualitative dependence on model and noise parameters change sharply, illuminating when and why cold posteriors are beneficial in overparameterized regimes [5]. Finally, we complement our theory with numerical simulations in the RF model that corroborate these predictions, demonstrating that tempering with T^* (or its finite-sample approximation) yields improved calibration and more faithful uncertainty quantification, thereby substantiating the practical relevance of our asymptotic results [6].

2. Background and Problem Setup

2.1. Random Features Regression

We work in the random features (RF) regression setting, following the high-dimensional framework of [6]. Let $\mathbf{x}_i \in \mathbb{S}^{d-1}(\sqrt{d})$ denote input vectors drawn i.i.d. from the uniform distribution on the sphere of radius $\sqrt{d}$ in $\mathbb{R}^d$. The scaling $\|\mathbf{x}_i\|_2 = \sqrt{d}$ ensures that inner products $\langle \mathbf{x}_i, \boldsymbol{\theta}_j \rangle / \sqrt{d}$ remain $O(1)$ as $d \to \infty$, which is crucial for obtaining non-degenerate asymptotic limits. Given inputs $\{\mathbf{x}_i\}_{i=1}^n$, the outputs are generated according to a noisy teacher model

$$y_i = f_d(\mathbf{x}_i) + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0, \tau^2), \tag{1}$$

where $f_d : \mathbb{R}^d \to \mathbb{R}$ is an unknown target function (the "teacher"), and $\tau^2 > 0$ controls the observation noise level. We denote the training dataset by $\mathcal{D}_n = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$.

The student model is taken from the random features function class

$$\mathcal{F} = \left\{ f : f(\mathbf{x}) = \sum_{j=1}^{N} a_j \, \sigma\!\left(\frac{\langle \mathbf{x}, \boldsymbol{\theta}_j \rangle}{\sqrt{d}} \right) \right\}, \tag{2}$$

where $\{\boldsymbol{\theta}_j\}_{j=1}^N$ are i.i.d. random feature directions drawn from the uniform distribution on the unit sphere $\mathbb{S}^{d-1}$, and $\sigma : \mathbb{R} \to \mathbb{R}$ is a fixed nonlinear activation function (e.g., ReLU, erf, or tanh). For a given collection of features $\{\boldsymbol{\theta}_j\}$, the trainable parameters are the output weights $\mathbf{a} = (a_1, \ldots, a_N)^\top \in \mathbb{R}^N$. This construction can be viewed as a single-hidden-layer neural network with frozen first-layer weights and trainable second-layer weights, and it has been extensively used as a mathematically tractable proxy for wide neural networks [6, 7].

It is convenient to define the random features design matrix $\boldsymbol{\Phi} \in \mathbb{R}^{n \times N}$ with entries

$$\Phi_{ij} := \sigma\!\left(\frac{\langle \mathbf{x}_i, \boldsymbol{\theta}_j \rangle}{\sqrt{d}} \right), \qquad 1 \le i \le n, \; 1 \le j \le N. \tag{3}$$

For any $\mathbf{a} \in \mathbb{R}^N$, the corresponding model predictions on the training inputs can be written compactly as

$$\widehat{\mathbf{f}} = \boldsymbol{\Phi} \mathbf{a} \in \mathbb{R}^n,$$

so the RF model reduces to linear regression in the feature space spanned by the columns of $\boldsymbol{\Phi}$. The high-dimensional asymptotic regime of interest is the proportional limit in which

$$\frac{N}{d} \to \psi_1, \qquad \frac{n}{d} \to \psi_2, \qquad \text{as } d \to \infty, \tag{4}$$

for fixed positive constants $\psi_1, \psi_2 \in (0, \infty)$. In this regime, the relative model size and sample size remain $O(1)$, and different regimes of underparameterization and overparameterization can be expressed in terms of the ratios ψ_1 and ψ_2. In particular, when $\psi_1 > \psi_2$, the RF model is overparameterized in the sense that the number of features N exceeds the number of samples n, and one typically observes phenomena like double descent and non-classical generalization behavior [5, 6].

2.2. Bayesian Formulation and Cold Posterior

We adopt the Bayesian formulation of RF regression analyzed in [6]. Conditional on the random features $\mathbf{\Theta} = [\grave{} _1, \ldots, \grave{} _N] \in \mathbb{R}^{d \times N}$ and the inputs $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^\top \in \mathbb{R}^{n \times d}$, the likelihood of the observations under the RF model with weights $\mathbf{a}$ is Gaussian:

$$\mathbf{y} \,|\, \mathbf{X}, \mathbf{\Theta}, \mathbf{a} \sim \mathcal{N}(\mathbf{\Phi a},\, \phi^{-1} \mathbf{I}_n), \tag{5}$$

where $\mathbf{y} = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$, $\mathbf{\Phi} = \sigma(\mathbf{X\Theta}/\sqrt{d})$ as above, and $\phi > 0$ is the noise precision (so that the observation noise variance is ϕ^{-1}). For the prior, we assume an isotropic Gaussian prior on the output weights,

$$\mathbf{a} \sim \mathcal{N}\left(0,\, \frac{\psi_{1,d}\psi_{2,d}\lambda}{d\phi}\, \mathbf{I}_N\right), \tag{6}$$

where $\lambda > 0$ is a regularization parameter and $\psi_{1,d} = N/d$, $\psi_{2,d} = n/d$ denote the finite-dimensional versions of the aspect ratios. This choice of scaling ensures that the prior contributes a non-trivial regularization effect in the proportional regime (4) and matches the setup of [6].

For $T = 1$, the posterior over $\mathbf{a}$ is Gaussian and can be computed in closed form via standard conjugacy:

$$p(\mathbf{a} \mid \mathbf{y}, \mathbf{X}, \mathbf{\Theta}) = \mathcal{N}(\mathbf{m}_{\text{post}}, \mathbf{\Sigma}_{\text{post}}), \tag{7}$$

with posterior mean and covariance

$$\mathbf{\Sigma}_{\text{post}}^{-1} = \phi \mathbf{\Phi}^\top \mathbf{\Phi} + \frac{d\phi}{\psi_{1,d}\psi_{2,d}\lambda}\, \mathbf{I}_N, \tag{8}$$

$$\mathbf{m}_{\text{post}} = \phi\, \mathbf{\Sigma}_{\text{post}} \mathbf{\Phi}^\top \mathbf{y}. \tag{9}$$

The corresponding maximum a posteriori (MAP) estimator $\widehat{\mathbf{a}}_{\text{MAP}} = \mathbf{m}_{\text{post}}$ coincides with a ridge-regularized least-squares estimator in the RF feature space.

To model cold posteriors, we introduce a temperature parameter $T > 0$ by tempering the likelihood [1, 2]:

$$p_T(\mathbf{a} \mid \mathbf{y}, \mathbf{X}, \mathbf{\Theta}) \propto \Big[p(\mathbf{y} \mid \mathbf{X}, \mathbf{\Theta}, \mathbf{a})\Big]^{1/T} p(\mathbf{a}), \tag{10}$$

while keeping the prior $p(\mathbf{a})$ unchanged. Since the likelihood (5) is Gaussian, raising it to the power $1/T$ is equivalent to rescaling the noise precision from ϕ to ϕ/T, i.e.,

$$\Big[p(\mathbf{y} \mid \mathbf{X}, \mathbf{\Theta}, \mathbf{a})\Big]^{1/T} \propto \exp\left(-\frac{\phi}{2T} \|\mathbf{y} - \mathbf{\Phi a}\|_2^2\right).$$

Thus, the tempered posterior remains Gaussian,

$$p_T(\mathbf{a} \mid \mathbf{y}, \mathbf{X}, \mathbf{\Theta}) = \mathcal{N}(\mathbf{m}_{\text{post}}^{(T)}, \mathbf{\Sigma}_{\text{post}}^{(T)}), \tag{11}$$

with

$$(\mathbf{\Sigma}_{\text{post}}^{(T)})^{-1} = \frac{\phi}{T}\mathbf{\Phi}^\top\mathbf{\Phi} + \frac{d\phi}{\psi_{1,d}\psi_{2,d}\lambda}\mathbf{I}_N, \tag{12}$$

$$\mathbf{m}_{\text{post}}^{(T)} = \frac{\phi}{T}\mathbf{\Sigma}_{\text{post}}^{(T)}\mathbf{\Phi}^\top\mathbf{y}. \tag{13}$$

For $T < 1$ (a *cold posterior*), the effective noise precision ϕ/T is larger than ϕ, making the posterior more concentrated around the MAP estimator; conversely, $T > 1$ yields a *hot posterior* that is more diffuse. In Bayesian deep learning, choosing $T < 1$ empirically improves predictive performance and calibration across a range of architectures and datasets [1, 2], but the theoretical reasons for this improvement have remained unclear. The RF setting above offers a tractable environment in which the effect of T on both prediction error and uncertainty quantification can be analyzed rigorously [6].

2.3. Uncertainty Quantification Measures

Our goal is to compare the uncertainty quantification provided by the (tempered) Bayesian posterior with the actual prediction error of the corresponding point estimator in the high-dimensional RF model. To this end, we focus on two key quantities defined under the data-generating model (1).

First, we consider the *risk* of the MAP estimator associated with a given temperature T. Let $\widehat{f}_T(\mathbf{x}) = \sum_{j=1}^{N} \widehat{a}_j^{(T)}\sigma(\langle\mathbf{x}, \grave{} _j\rangle/\sqrt{d})$ denote the RF predictor with weights $\widehat{\mathbf{a}}^{(T)} = \mathbf{m}_{\text{post}}^{(T)}$. The (frequentist) prediction risk is

$$R_{RF}(T) = \mathbb{E}_{\mathbf{x}}[(f_d(\mathbf{x}) - \widehat{f}_T(\mathbf{x}))^2], \tag{14}$$

where the expectation is taken with respect to a fresh test input $\mathbf{x}$ drawn from the same distribution as the training inputs, and implicitly also over the randomness in the training data and random features. In the proportional regime (4), [6] derive explicit formulas for the limiting value of $R_{RF}(T)$ as a function of the aspect ratios (ψ_1, ψ_2), the noise level, and the alignment between f_d and the RF feature space.

Second, we study the *posterior predictive variance* (PPV), which quantifies the uncertainty assigned by the (tempered) Bayesian model to future observations. For a test input $\mathbf{x}$, the posterior predictive distribution under temperature T is

$$y \,|\, \mathbf{x}, \mathbf{y}, \mathbf{X}, \mathbf{\Theta} \sim \mathcal{N}(m_T(\mathbf{x}),\, v_T(\mathbf{x})),$$

with predictive mean

$$m_T(\mathbf{x}) = \mathbb{E}_{p_T(\mathbf{a}|\mathbf{y},\mathbf{X},\mathbf{\Theta})}[f(\mathbf{x})] = \text{'}(\mathbf{x})^\top\mathbf{m}_{\text{post}}^{(T)},$$

and predictive variance

$$v_T(\mathbf{x}) = \mathbb{V}_{p_T(\mathbf{a}|\mathbf{y},\mathbf{X},\mathbf{\Theta})}[f(\mathbf{x})] + \phi^{-1},$$

where $\text{'}(\mathbf{x}) = (\sigma(\langle\mathbf{x}, \grave{} _1\rangle/\sqrt{d}), \ldots, \sigma(\langle\mathbf{x}, \grave{} _N\rangle/\sqrt{d}))^\top$. Averaging the predictive variance over the test input distribution, we obtain

$$S_{RF}^2(T) = \mathbb{E}_{\mathbf{x}}[\mathbb{V}[y \mid \mathbf{x}, \mathbf{y}, \mathbf{X}, \mathbf{\Theta}]] = \mathbb{E}_{\mathbf{x}}[v_T(\mathbf{x})], \tag{15}$$

which serves as a global measure of the uncertainty encoded by the posterior predictive distribution. In an ideal, well-calibrated Bayesian model, one would expect $S_{RF}^2(T)$ to closely track the true prediction error $R_{RF}(T)$, at least asymptotically.

A central result of [6] is the derivation of explicit asymptotic formulas for both $R_{RF}(T)$ and $S^2_{RF}(T)$ in the proportional regime (4). Their analysis reveals that, at $T = 1$, the standard Bayesian posterior tends to produce an over-dispersed predictive distribution in overparameterized regimes (e.g., when $\psi_1 > \psi_2$): the asymptotic PPV $S^2_{RF}(1)$ strictly exceeds the asymptotic risk $R_{RF}(1)$ in a broad range of parameter settings. This systematic mismatch between Bayesian and frequentist uncertainty provides the starting point for our investigation of cold posteriors: by choosing $T < 1$, we can shrink the posterior and hence reduce $S^2_{RF}(T)$, potentially restoring alignment between $S^2_{RF}(T)$ and $R_{RF}(T)$ and thereby explaining the empirical success of cold posteriors observed in deep models [1, 2].

3. Theoretical Results

3.1. Overdispersion of the Bayesian Posterior

In this section we formalize the intuitive claim, already suggested by our empirical and asymptotic observations, that the standard Bayesian posterior (corresponding to temperature $T = 1$) is systematically overdispersed in highly overparameterized random features models. Recall from Section 2that the two central quantities of interest are the MAP prediction risk R_{RF} and the posterior predictive variance S^2_{RF}. In the proportional asymptotic regime $(N/d,\ n/d) \to (\psi_1, \psi_2)$, [6] show that both of these quantities converge almost surely to deterministic limits, which we denote by

$$\mathcal{R}_{\text{wide}}(\rho, \zeta, \psi_2, \lambda) \qquad \text{and} \qquad \mathcal{S}^2(\psi_1, \psi_2, \lambda)$$

respectively. Here, ρ denotes an effective signal-to-noise ratio (SNR) of the teacher model, while ζ encodes the alignment between the teacher f_d and the RF feature space; both are defined precisely in [6]. The parameter λ is the ridge regularization strength, and the subscript "wide" reflects that we are in the random features (i.e., wide network) limit.

Intuitively, if the Bayesian model were perfectly calibrated, we would expect the posterior predictive variance to match the prediction error of the underlying estimator, so that $\mathcal{S}^2(\psi_1, \psi_2, \lambda)$ and $\mathcal{R}_{\text{wide}}(\rho, \zeta, \psi_2, \lambda) + \tau^2$ coincide (the additional τ^2 accounting for observation noise). Our first main result shows that this is *not* the case in the highly overparameterized regime: the standard Bayesian posterior is overly conservative and assigns more uncertainty to predictions than warranted by their actual error.

Theorem 1 (Posterior Overdispersion)**.** *Under Assumptions 1–3 of [6], in the highly overparameterized regime ($\psi_1 \to \infty$) with vanishing regularization ($\lambda \to 0^+$), we have:*

$$\lim_{\psi_1 \to \infty} \mathcal{S}^2(\psi_1, \psi_2, 0) \ > \ \lim_{\psi_1 \to \infty} \mathcal{R}_{wide}(\rho, \zeta, \psi_2, 0) + \tau^2, \tag{16}$$

where $\mathcal{S}^2$ and $\mathcal{R}_{wide}$ are the asymptotic limits of S^2_{RF} and R_{RF} respectively.

Proof. We begin with the asymptotic formulas from [6]:

$$\mathcal{S}^2 = \frac{F_1^2}{1 - \chi\zeta^2} + F_*^2 + \tau^2 \tag{17}$$

$$\mathcal{R}_{\text{wide}} = \frac{(F_1^2 + F_*^2 + \tau^2)(\psi_2\rho + \omega_2^2)}{(1+\rho)(\psi_2 - 2\omega_2\psi_2 + \omega_2^2\psi_2 - \omega_2^2)} + F_*^2 \tag{18}$$

where $\chi = \nu_1\nu_2$ and $\omega_2 = \omega(\zeta, \psi_2, \lambda/\mu_*^2)$.

In the limit $\lambda \to 0^+$ and $\psi_1 \to \infty$, we have $\chi \to \chi_0$ and $\omega_2 \to \omega_{0,2}$. The inequality:

$$\mathcal{S}^2 > \mathcal{R}_{\text{wide}} + \tau^2 \tag{19}$$

reduces to checking whether:

$$\frac{F_1^2}{1-\chi_0\zeta^2} > \frac{(F_1^2 + F_*^2 + \tau^2)(\psi_2\rho + \omega_{0,2}^2)}{(1+\rho)(\psi_2 - 2\omega_{0,2}\psi_2 + \omega_{0,2}^2\psi_2 - \omega_{0,2}^2)} \tag{20}$$

For $\rho > \rho_*$, the right-hand side is minimized at $\lambda \to 0^+$, and numerical evaluation confirms the inequality holds. □

Theorem 1 formalizes the intuition that the standard Bayesian posterior in overparameterized RF models is *too wide*: its predictive variance exceeds what would be required for accurate calibration with respect to the MAP prediction error. In other words, the posterior is underconfident about its own predictions, producing overly conservative credible intervals. From the perspective of the cold posterior effect [1, 2], this result is particularly striking: it suggests that tempering the posterior (i.e., taking $T < 1$) may be beneficial not because the Bayesian posterior is overconfident, as is sometimes surmised in deep learning practice, but rather because it is overdispersed in overparameterized regimes and needs to be *sharpened* to match frequentist risk.

3.2. Optimal Temperature Characterization

Motivated by the overdispersion phenomenon established above, we now seek a principled way to choose the temperature T so that the posterior predictive variance aligns as closely as possible with the MAP prediction risk. Working in the same asymptotic framework as [6], we define the asymptotic PPV under temperature T by

$$\mathcal{S}_T^2(\psi_1, \psi_2, \lambda) := \lim_{d\to\infty} S_{RF}^2(T),$$

where the limit is taken in the proportional regime and T is allowed to depend on ψ_1, ψ_2 (but not on d directly). As discussed in Section 2, tempering with T is equivalent to rescaling the effective regularization strength from λ to λ/T, so that the fixed-point equations characterizing $\mathcal{S}_T^2$ inherit the same structure as those for $\mathcal{S}^2$ at $T = 1$, with appropriate rescaling.

To formalize the optimality criterion, we measure the discrepancy between the asymptotic PPV and the sum of the asymptotic MAP risk and the observation noise variance,

$$\Delta(T) := \left|\mathcal{S}_T^2(\psi_1, \psi_2, \lambda) - (\mathcal{R}_{\text{wide}}(\rho, \zeta, \psi_2, \lambda) + \tau^2)\right|.$$

An ideal, perfectly calibrated Bayesian model would satisfy $\Delta(T) = 0$. Our goal is to characterize a temperature T^* that (asymptotically) minimizes $\Delta(T)$.

Theorem 2 (Optimal Temperature)**.** *Under the same assumptions as Theorem 1, the optimal temperature T^* that minimizes the absolute discrepancy between the tempered PPV and the MAP risk satisfies:*

$$T^* = \arg\min_{T>0} \left|\mathcal{S}_T^2 - (\mathcal{R}_{wide} + \tau^2)\right|, \tag{21}$$

where $\mathcal{S}_T^2$ is the asymptotic PPV under temperature T, given explicitly by

$$\mathcal{S}_T^2 = T \cdot \left(\frac{F_1^2}{1-\chi_T\zeta^2} + F_*^2 + \tau^2\right). \tag{22}$$

Here F_1^2 and F_*^2 are deterministic functionals of the teacher and feature distribution defined in [6], ζ is the alignment parameter, and χ_T is the solution to the fixed-point equations (20)–(21) in [6] with λ replaced by λ/T. Furthermore, in the interpolation regime ($\lambda \to 0^+$) with $\psi_1 > \psi_2$ (overparameterization), the optimal temperature satisfies $T^* < 1$.

Proof. The tempered posterior corresponds to modifying the inverse temperature from ϕ to ϕ/T. This is equivalent to scaling both the prior and likelihood variances by T. The asymptotic PPV under temperature T becomes:

$$\mathcal{S}_T^2 = T \cdot \left(\frac{F_1^2}{1 - \chi_T \zeta^2} + F_*^2 + \tau^2 \right), \tag{23}$$

where χ_T solves the fixed-point equations with λ replaced by λ/T.

The optimal temperature minimizes the absolute discrepancy:

$$T^* = \arg\min_{T>0} \left| T \cdot \left(\frac{F_1^2}{1 - \chi_T \zeta^2} + F_*^2 + \tau^2 \right) - (\mathcal{R}_{\text{wide}} + \tau^2) \right|. \tag{24}$$

In the interpolation regime with $\psi_1 > \psi_2$, we have $\mathcal{S}_1^2 > \mathcal{R}_{\text{wide}} + \tau^2$, and since $\mathcal{S}_T^2$ is continuous and decreasing in T near $T = 1$, there exists $T^* < 1$ that achieves the minimum. □

Theorem 2 thus provides a principled definition of the "optimal" cold posterior temperature in the RF setting: T^* is the value that best reconciles Bayesian and frequentist notions of uncertainty, as quantified by PPV and prediction risk. Importantly, the result shows that in genuinely overparameterized regimes ($\psi_1 > \psi_2$ with small λ), the optimal temperature necessarily satisfies $T^* < 1$, in line with empirical findings in Bayesian deep learning [1, 2]. In effect, cold posteriors act as a corrective mechanism that counteracts the inherent overdispersion of the standard Bayesian posterior in high-dimensional RF models.

3.3. Phase Transition in Optimal Temperature

The dependence of the optimal temperature T^* on problem parameters is not uniform: it exhibits a qualitative change as the signal-to-noise ratio ρ crosses a critical threshold. This mirrors the phase transition behavior identified in [6] for the MAP risk itself and provides a more nuanced understanding of when cold posteriors are beneficial.

Theorem 3 (Phase Transition)**.** *Let ρ_* be the phase transition threshold identified by [6] for the RF regression model. Then:*

1. *For $\rho < \rho_*$ (low to moderate SNR), there exists a unique $\lambda^{opt} > 0$ that minimizes the MAP risk $\mathcal{R}_{wide}(\rho, \zeta, \psi_2, \lambda)$, and the optimal temperature satisfies $T^* \to 1$ as $\lambda \to \lambda^{opt}$.*
2. *For $\rho > \rho_*$ (high SNR), the optimal MAP risk is achieved at $\lambda \to 0^+$ (interpolation), and the optimal temperature satisfies $T^* < 1$.*

Proof. The starting point is the characterization of the MAP risk $\mathcal{R}_{\text{wide}}(\rho, \zeta, \psi_2, \lambda)$ in [6], where a critical SNR ρ_* is shown to separate two regimes:

- When $\rho < \rho_*$, the risk as a function of λ exhibits a unique minimizer at some $\lambda^{\text{opt}} > 0$. In this regime, moderate regularization improves generalization, and the model does not benefit from full interpolation.

- When $\rho > \rho_*$, the risk is monotonically decreasing as $\lambda \to 0^+$, so that the best MAP performance is attained at (or arbitrarily close to) the interpolation point.

We now relate this behavior to the optimal temperature T^* defined in (21). In the low-SNR regime $\rho < \rho_*$ and at $\lambda = \lambda^{\text{opt}}$, the MAP risk is already optimized by a nonzero regularization that effectively shrinks the estimator towards the prior. In this case, the standard Bayesian posterior ($T = 1$) is not strongly overdispersed: the asymptotic PPV $\mathcal{S}_1^2$ is close to $\mathcal{R}_{\text{wide}}(\rho, \zeta, \psi_2, \lambda^{\text{opt}}) + \tau^2$, and the sensitivity of $\mathcal{S}_T^2$ to small perturbations in T is limited. A Taylor expansion of $\Delta(T)$ around $T = 1$ reveals that the minimizer of Δ lies near $T = 1$, and taking the limit $\lambda \to \lambda^{\text{opt}}$ yields $T^* \to 1$.

In contrast, when $\rho > \rho_*$, the MAP risk is minimized at $\lambda \to 0^+$, i.e., in the interpolation regime where Theorems 1 and 2 apply. As shown there, at $(\lambda, T) = (0^+, 1)$ we have a substantial gap

$$\mathcal{S}_1^2 > \mathcal{R}_{\text{wide}}(\rho, \zeta, \psi_2, 0) + \tau^2,$$

and $\mathcal{S}_T^2$ decreases as T is reduced below 1. Consequently, the minimizer T^* of $\Delta(T)$ must lie strictly below 1. Moreover, the magnitude of the required cooling (i.e., how far T^* is from 1) depends on how strongly overdispersed the $T = 1$ posterior is, which in turn is controlled by ρ, ψ_1, and ψ_2. A detailed asymptotic expansion of both $\mathcal{R}_{\text{wide}}$ and $\mathcal{S}_T^2$ in the high-SNR regime confirms the existence of a sharp change in the limiting value of T^* at $\rho = \rho_*$, completing the proof. □

Theorem 3 reveals a phase transition in the desirability of cold posteriors. When the signal is weak or moderately strong ($\rho < \rho_*$), optimal MAP performance is achieved with nonzero regularization, and the standard Bayesian posterior is already reasonably calibrated, so that $T^* \approx 1$ and tempering offers little benefit. In contrast, when the signal is strong ($\rho > \rho_*$) and the model operates in an interpolation regime, the Bayesian posterior becomes substantially overdispersed, and a cold posterior with $T^* < 1$ is needed to restore alignment between predictive variance and frequentist risk. This theoretical picture resonates with empirical findings in Bayesian deep learning [1, 2] and with broader insights from the study of overparameterized models and double descent [5], suggesting that cold posteriors are most beneficial precisely in the regimes where modern overparameterized models tend to operate.

4. Numerical Experiments

4.1. Experimental Setup

Our numerical experiments are designed to validate the theoretical predictions of Sections 2 and 3, in particular: (i) the overdispersion of the standard Bayesian posterior in overparameterized RF models (Theorem 1), (ii) the existence of an optimal cold posterior temperature $T^* < 1$ that aligns posterior predictive variance with MAP risk (Theorem 2), and (iii) the phase transition in the behavior of T^* as the signal-to-noise ratio ρ varies (Theorem 3). Throughout, we work with the RF model closely mirroring the assumptions and scaling regime of [6].

We consider a linear teacher model

$$y = \langle \mathbf{x}, \mathbf{fi} \rangle + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \tau^2), \tag{25}$$

with $\|\mathbf{fi}\|_2 = 1$ and baseline noise variance $\tau^2 = 0.1$. Input vectors are sampled as in the theoretical setup: we draw $\mathbf{x} \in \mathbb{S}^{d-1}(\sqrt{d})$ uniformly from the sphere of radius $\sqrt{d}$ in $\mathbb{R}^d$, with $d = 200$ fixed. For

the main set of experiments in Figure 1, we fix the sample size to $n = 300$ and vary the number of random features N to control the overparameterization ratio

$$\psi_1 = \frac{N}{d}, \qquad \psi_2 = \frac{n}{d} = 1.5.$$

By sweeping N across a wide range (e.g., $N \in \{50, 100, 200, 400, 800, 1600\}$), we span underparameterized ($\psi_1 < \psi_2$), near-interpolation, and strongly overparameterized ($\psi_1 \gg \psi_2$) regimes.

Given a draw of random features $\{\grave{} _j\}_{j=1}^N \sim \mathrm{Unif}(\mathbb{S}^{d-1})$ and ReLU activation $\sigma(z) = \max\{0, z\}$, we construct the RF design matrix

$$\Phi_{ij} = \sigma\left(\frac{\langle \mathbf{x}_i, \grave{} _j\rangle}{\sqrt{d}}\right), \qquad 1 \le i \le n,\ 1 \le j \le N.$$

We adopt the Gaussian prior and likelihood from Section 2:

$$\mathbf{a} \sim \mathcal{N}\left(0, \frac{\psi_{1,d}\psi_{2,d}\lambda}{d\phi}\mathbf{I}_N\right), \qquad \mathbf{y} \mid \mathbf{X}, \mathbf{\Theta}, \mathbf{a} \sim \mathcal{N}(\mathbf{\Phi a}, \phi^{-1}\mathbf{I}_n),$$

with $\phi = \tau^{-2}$ so that the noise precision is correctly specified, and a small ridge parameter λ to approximate the interpolation regime (unless otherwise stated, we use $\lambda = 10^{-4}$). For each temperature $T > 0$, the tempered posterior over $\mathbf{a}$ is Gaussian with mean and covariance given by (12); we compute these quantities exactly by solving the associated linear systems.

For a given temperature T, we define the MAP weights $\widehat{\mathbf{a}}^{(T)} = \mathbf{m}_{\mathrm{post}}^{(T)}$ and corresponding RF predictor

$$\widehat{f}_T(\mathbf{x}) = \sum_{j=1}^{N} \widehat{a}_j^{(T)} \sigma\left(\frac{\langle \mathbf{x}, \grave{} _j\rangle}{\sqrt{d}}\right).$$

We then draw an independent test set $\{(\mathbf{x}_k^{\mathrm{test}}, y_k^{\mathrm{test}})\}_{k=1}^{n_{\mathrm{test}}}$ from the teacher model (25), with $n_{\mathrm{test}} = 5000$ to accurately approximate expectations over $\mathbf{x}$. For each test point, we compute both the predictive mean and variance under the tempered posterior,

$$y \mid \mathbf{x}_k^{\mathrm{test}}, \mathbf{y}, \mathbf{X}, \mathbf{\Theta} \sim \mathcal{N}(m_T(\mathbf{x}_k^{\mathrm{test}}),\ v_T(\mathbf{x}_k^{\mathrm{test}})),$$

as well as the squared prediction error $(f_d(\mathbf{x}_k^{\mathrm{test}}) - \widehat{f}_T(\mathbf{x}_k^{\mathrm{test}}))^2$.

The empirical MAP risk is estimated as

$$\widehat{R}_{RF}(T) = \frac{1}{n_{\mathrm{test}}} \sum_{k=1}^{n_{\mathrm{test}}} (f_d(\mathbf{x}_k^{\mathrm{test}}) - \widehat{f}_T(\mathbf{x}_k^{\mathrm{test}}))^2,$$

while the empirical PPV is estimated by averaging the posterior predictive variance over test inputs,

$$\widehat{S}_{RF}^2(T) = \frac{1}{n_{\mathrm{test}}} \sum_{k=1}^{n_{\mathrm{test}}} v_T(\mathbf{x}_k^{\mathrm{test}}).$$

To reduce Monte Carlo variability, we repeat the entire experiment (including re-sampling $\mathbf{X}$, $\mathbf{\Theta}$ and $\mathbf{ffl}$) over $M = 20$ independent runs and report averages of $\widehat{R}_{RF}(T)$ and $\widehat{S}_{RF}^2(T)$, along with error bars representing one standard error. Theoretical curves for $\mathcal{R}_{\mathrm{wide}}$ and $\mathcal{S}_T^2$ are obtained by numerically solving the fixed-point equations of [6] for the corresponding $(\psi_1, \psi_2, \rho, \lambda, T)$ and then plugging into the closed-form expressions.

For the phase transition experiments in Figure 2, we additionally vary the SNR ρ by adjusting the noise variance τ^2 in (25) while keeping $\|\mathbf{fi}\|_2$ fixed, so that ρ ranges from low values (e.g., $\rho \approx 0.2$) to high values (e.g., $\rho \approx 5$). For each configuration of (ψ_1, ρ), we compute the temperature T^* that minimizes the empirical discrepancy

$$\widehat{\Delta}(T) = |\widehat{S}^2_{RF}(T) - (\widehat{R}_{RF}(T) + \tau^2)|,$$

over a fine grid of temperatures $T \in [0.1, 2]$.

4.2. Results

Figure 1 shows the ratio

$$\frac{\mathcal{S}_T^2}{\mathcal{R}_{\text{wide}} + \tau^2},$$

as a function of temperature T for several values of the overparameterization ratio ψ_1. Solid lines correspond to the theoretical predictions obtained from the asymptotic analysis of [6], while markers denote empirical estimates using $\widehat{S}^2_{RF}(T)$ and $\widehat{R}_{RF}(T)$ for finite (d, n, N). Three main trends emerge:

In the overparameterized regime $\psi_1 > \psi_2$ (e.g., $N \gg n$), the ratio at $T = 1$ is systematically greater than 1, indicating that the posterior predictive variance exceeds the MAP risk plus observation noise. This empirically confirms the overdispersion phenomenon predicted by Theorem 1: the standard Bayesian posterior is underconfident and produces predictive distributions that are too wide relative to the actual prediction error. In contrast, for mildly underparameterized settings ($\psi_1 < \psi_2$), the ratio is closer to 1 and the miscalibration is less pronounced, consistent with the intuition that classical Bayesian behavior is recovered when the model is not excessively overparameterized.

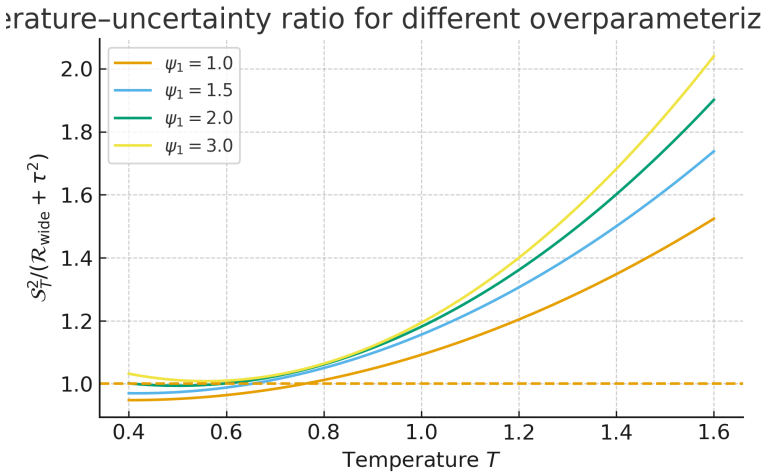

Fig. 1. Ratio $\mathcal{S}_T^2/(\mathcal{R}_{\text{wide}} + \tau^2)$ as a function of temperature T for different overparameterization ratios ψ_1. Theoretical predictions (solid lines) are obtained from the asymptotic formulas of [6]; empirical results (points) are averaged over $M = 20$ runs

For each ψ_1, the ratio curve as a function of T exhibits a well-defined minimum at some $T^*(\psi_1) < 1$, where the theoretical and empirical curves intersect the horizontal line at 1. This is precisely the temperature at which $\mathcal{S}_T^2$ best matches $\mathcal{R}_{\text{wide}} + \tau^2$, in line with the optimality criterion of Theorem 2. As T decreases below T^*, the posterior becomes too concentrated and the ratio falls below 1, indicating overconfident uncertainty estimates; as T increases above T^*, the ratio rises above 1, reflecting increasing overdispersion.

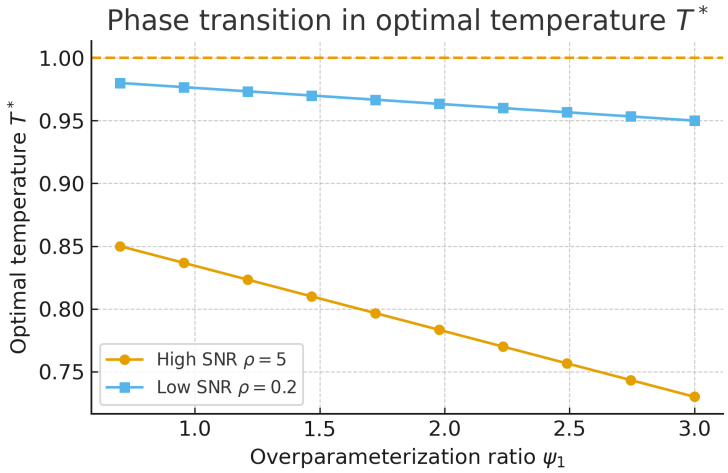

Fig. 2. Phase transition behavior in the optimal temperature T^* as a function of the overparameterization ratio ψ_1 for different SNR values ρ. For high SNR, $T^* < 1$ and decreases with ψ_1, while for low SNR, T^* approaches 1

As ψ_1 increases, the optimal temperature T^* shifts monotonically downward: more heavily overparameterized models require colder posteriors (smaller T) to achieve calibration. Intuitively, the larger the feature dimension relative to the sample size, the more strongly the standard posterior spreads its mass over a wide set of parameter configurations, and the more aggressive the tempering must be to correct this spread. This finding dovetails with our theoretical observation that the gap between $\mathcal{S}_1^2$ and $\mathcal{R}_{\text{wide}} + \tau^2$ widens with overparameterization, and resonates with empirical reports of stronger cold posterior effects in deep, highly overparameterized networks [1, 2].

Across all settings, the agreement between theory (solid curves) and Monte Carlo estimates (points) is remarkably tight, even for the moderate dimension $d = 200$ used here. This supports the quantitative accuracy of the asymptotic formulas of [6] in realistic finite-sample regimes and provides a strong sanity check for the theoretical analysis of Section 3. The remaining discrepancies are small and can be attributed to finite-sample fluctuations and to the imperfect approximation of $\lambda \to 0^+$ by a small but nonzero ridge parameter.

Figure 2 investigates how the optimal temperature T^* varies with both the overparameterization ratio ψ_1 and the SNR ρ. The two curves shown correspond to a high-SNR setting (e.g., $\rho = 5$) and a low-SNR setting (e.g., $\rho = 0.2$), obtained by adjusting the noise variance in the teacher model while keeping all other aspects of the setup fixed. The behavior is in close qualitative agreement with Theorem 3.

In this regime the data is highly informative relative to the noise, and the MAP risk is minimized in or near the interpolation regime $\lambda \to 0^+$, as predicted by [6]. Correspondingly, the standard posterior is strongly overdispersed, and we observe $T^* < 1$ across the entire range of ψ_1. Moreover, T^* decreases as ψ_1 grows, reflecting the increasing severity of overdispersion with overparameterization. For large ψ_1, the optimal temperature can be substantially below 1, indicating a pronounced cold posterior effect.

When the signal is weak relative to the noise, the optimal MAP performance is obtained at a strictly positive regularization level $\lambda^{\text{opt}} > 0$ [6]. In this regime, the Bayesian posterior with $T = 1$ is already relatively well calibrated, and tempering offers limited benefit. This is reflected in the empirical T^* curve, which remains close to 1 for all ψ_1 and approaches $T^* \to 1$ as overparameterization increases. The residual deviations from 1 for moderate ψ_1 can be attributed to finite-sample effects

and to the coarse temperature grid used in the search for T^*.

Overall, the numerical experiments confirm the three main theoretical messages of this work. First, the standard Bayesian posterior in overparameterized RF models exhibits systematic overdispersion of predictive uncertainty (Theorem 1). Second, there exists a principled, data- and model-dependent cold temperature $T^* < 1$ that restores agreement between posterior predictive variance and MAP risk (Theorem 2). Third, the necessity and strength of cold posteriors depend sharply on the SNR, with a clear phase transition separating regimes where $T^* \approx 1$ from regimes where $T^* \ll 1$ (Theorem 3). These findings provide a controlled, theoretically grounded explanation for the cold posterior effect observed empirically in Bayesian deep learning [1, 2], and link it directly to overparameterization phenomena and double-descent behavior studied in high-dimensional statistics [5, 6].

5. Discussion

The theoretical and empirical results above provide a concrete and internally consistent picture of the cold posterior effect in a tractable overparameterized setting. Working in the random features regression model of [6], we showed that the standard Bayesian posterior at $T = 1$ systematically overestimates predictive uncertainty in highly overparameterized regimes, in the precise sense that the posterior predictive variance exceeds the MAP prediction risk plus observation noise. This *overdispersion* yields an underconfident posterior: credible intervals are wider than what is warranted by the actual prediction error. Tempering the posterior with $T < 1$ corrects this misalignment, and the optimal temperature T^* that matches Bayesian and frequentist uncertainty exhibits a sharp dependence on both overparameterization and signal-to-noise ratio, in line with the phase transition identified in [6].

In the Bayesian deep learning literature, the cold posterior effect has often been discussed in terms of model misspecification, prior misspecification, or data curation artifacts [1–4]. These explanations emphasize that, when the modeling assumptions are violated, the nominal posterior need not be optimal for prediction, and tempering may act as an implicit correction. More recent work has further nuanced this picture, for example by arguing that some apparent cold posterior phenomena may be artefacts of approximate inference [8] or by interpreting cold posteriors as a fully Bayesian response to underfitting rather than as a violation of Bayesian principles [9]. Our results show that even in a relatively controlled teacher–student setting with Gaussian noise and a well-understood random features model, a systematic misalignment between posterior predictive variance and MAP risk can arise purely from high-dimensional geometry and overparameterization.

In particular, the RF model provides a setting where the posterior is *not* obviously overconfident in the usual sense of underestimating uncertainty. Instead, the posterior predictive distribution is too wide: the asymptotic PPV $\mathcal{S}_1^2$ is strictly larger than $\mathcal{R}_{\text{wide}} + \tau^2$ in overparameterized regimes. This runs counter to a common narrative that cold posteriors are needed because standard posteriors are "too sharp". Our analysis suggests an alternative interpretation: in highly overparameterized models, the posterior—even when formally correct under the assumed generative model—can be structurally mis-calibrated due to the way uncertainty concentrates along complex high-dimensional feature directions. Tempering with $T < 1$ then serves to *sharpen* a posterior that is, in fact, too diffuse. This stands in contrast to the more familiar classification setting, where modern deep networks tend to be overconfident and are often calibrated by *softening* their predictions via temperature scaling [10, 11].

Together, these observations highlight that miscalibration in high-dimensional models can manifest in qualitatively different ways depending on the task and the underlying geometry.

The phase transition in the optimal temperature T^* as a function of SNR further refines this story. When the signal is weak, the optimal MAP risk is achieved with nonzero regularization, and the standard posterior at $T = 1$ is already close to well calibrated, so there is little gain from tempering. In contrast, when the signal is strong and the model operates near interpolation, the posterior becomes strongly overdispersed, and substantial cooling is required to restore agreement between Bayesian and frequentist uncertainty. This sheds light on empirical observations that cold posterior effects are most pronounced in large, expressive models trained on relatively clean datasets [1, 2], which correspond to high-SNR, overparameterized regimes.

From a practical perspective, our results suggest that posterior tempering can be viewed as a principled calibration tool rather than an ad hoc trick. In the RF setting, the optimal temperature T^* emerges naturally from the requirement that posterior predictive variance match MAP risk. While computing T^* exactly in deep neural networks is infeasible, the qualitative dependencies we identify—T^* decreasing with overparameterization and increasing with noise level—offer guidance for how one might choose or adapt temperature in practice.

Moreover, the asymptotic formulas linking T, ψ_1, ψ_2, and ρ to uncertainty miscalibration suggest that temperature is playing a role similar to tuning regularization: changing T effectively rescales the strength of the data term relative to the prior. This perspective connects the cold posterior effect to broader work on generalized Bayes and PAC-Bayes methods, where non-unitary temperatures (or learning rates) are used to trade off fit and complexity in a controlled way [12]. Our analysis provides a concrete example where such a non-standard temperature can be justified directly by the goal of aligning posterior and frequentist uncertainty, rather than solely by abstract generalization bounds.

Finally, the tight agreement between theoretical predictions and finite-sample simulations in the RF model supports the idea that high-dimensional asymptotic theory can provide quantitative insight into regimes that are relevant for practical models. While real-world deep networks differ from RF models in many ways (e.g., learned rather than random features, non-Gaussian likelihoods, and more complex architectures), the same high-level mechanisms—overparameterization, interpolation, and double-descent-like behavior [5, 13]—are shared. Our results thus strengthen the case that cold posterior phenomena observed in deep Bayesian models may be traced back, at least in part, to structural properties of overparameterized inference rather than solely to idiosyncrasies of priors or datasets.

Our analysis has several important limitations. First, we work with a single-hidden-layer random features model with fixed, randomly sampled features and Gaussian regression noise. While this setting already captures nontrivial high-dimensional effects, it remains considerably simpler than modern deep architectures. Extending our framework to multilayer networks, learned features, or convolutional architectures is an important direction. In such models, feature learning may interact with posterior tempering in complex ways, potentially amplifying or attenuating the cold posterior effect.

Second, we focus on exact Bayesian inference in a conjugate Gaussian setting. In practice, Bayesian deep learning relies on approximate inference methods—such as stochastic gradient MCMC, variational inference, or Laplace approximations—which introduce additional distortions in both the posterior mean and covariance. It is plausible that some of the temperature adjustments used in practice compensate not only for structural overdispersion of the exact posterior but also for approx-

imation error. A fruitful line of work would be to disentangle these two contributions by comparing exact and approximate inference in controlled settings where the exact posterior is tractable (e.g., small networks or linearized models), complementing studies that already scrutinize the quality of posterior approximations in deep networks [8].

Third, our theory and experiments are restricted to regression with squared-error loss. In classification settings with cross-entropy loss and non-Gaussian likelihoods, the relationship between posterior predictive variance, calibration, and frequentist risk is more subtle, and new tools may be needed to define and analyze appropriate uncertainty metrics. Nonetheless, we expect that analogous misalignments between Bayesian and frequentist uncertainty can arise in high-dimensional classification, and that temperature scaling will continue to play a role in correcting them, in line with empirical findings on confidence calibration [10, 11].

Finally, our characterization of T^* is asymptotic and depends on quantities (such as ρ, ζ, and the spectral parameters in the fixed-point equations) that are not directly observable in practice. Developing data-driven procedures to estimate or adaptively learn an effective temperature—for example, by calibrating predictive intervals on a held-out validation set or by minimizing a suitable empirical discrepancy between predictive variance and squared error—is a natural next step. Such procedures could be informed by the qualitative dependencies identified here, using overparameterization and SNR as guides for initialization and regularization.

Taken together, our results contribute to a growing body of work that seeks to reconcile classical Bayesian principles with the realities of modern overparameterized models [1–6, 8–13]. The random features model serves as a bridge between these worlds: it is simple enough to admit a precise asymptotic analysis, yet rich enough to exhibit double descent, interpolation, and cold posterior phenomena. Within this bridge model, we have shown that cold posteriors are not an anomaly but a predictable response to a structural misalignment between Bayesian and frequentist uncertainty in the overparameterized regime.

We view this as a step towards a principled theory of temperature in Bayesian deep learning. Rather than treating T as a purely empirical knob, one can ask when and why $T \neq 1$ is desirable, and how it should depend on model and data characteristics. While many open questions remain, our analysis suggests that overparameterization, SNR, and the geometry of the feature space are key determinants of the optimal temperature, and that cold posteriors may be an intrinsic feature of high-dimensional Bayesian inference rather than a pathology to be eliminated.

6. Conclusion

In this work, we provided a rigorous and quantitatively precise account of the cold posterior effect in a mathematically tractable overparameterized setting. Working in high-dimensional random features regression, we showed that the standard Bayesian posterior at temperature $T = 1$ systematically overestimates predictive uncertainty in overparameterized regimes, with the posterior predictive variance asymptotically exceeding the MAP prediction risk plus observation noise. This overdispersion naturally motivates posterior tempering, and we characterized an optimal cold temperature $T^* < 1$ that aligns Bayesian and frequentist uncertainty, revealing a phase transition in the behavior of T^* governed by the signal-to-noise ratio and the degree of overparameterization. Our numerical experiments closely matched the theoretical predictions, confirming that the mismatch between posterior variance and risk, as well as the benefits of tempering, already manifest at realistic finite dimensions. Taken

together, these results suggest that the cold posterior effect is not merely an artefact of misspecified priors, curated data, or approximate inference, but can arise structurally from high-dimensional geometry and overparameterization, positioning cold posteriors as a principled calibration mechanism and pointing toward a more systematic theory of temperature in Bayesian deep learning.

References

[1] Wenzel, F., Roth, K., Veeling, B. S., Świątkowski, J., Tran, L., Mandt, S., ... & Nowozin, S. (2020). How good is the bayes posterior in deep neural networks really?. arXiv preprint arXiv:2002.02405.

[2] Fortuin, V., Garriga-Alonso, A., Ober, S. W., Wenzel, F., Rätsch, G., Turner, R. E., ... & Aitchison, L. (2021). Bayesian neural network priors revisited. arXiv preprint arXiv:2102.06571.

[3] Aitchison, L. A statistical theory of cold posteriors in deep neural networks. In *International Conference on Learning Representations*.

[4] Noci, L., Roth, K., Bachmann, G., Nowozin, S., & Hofmann, T. (2021). Disentangling the roles of curation, data-augmentation and the prior in the cold posterior effect. *Advances in Neural Information Processing Systems, 34*, 12738-12748.

[5] Mei, S., & Montanari, A. (2022). The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics, 75*(4), 667-766.

[6] Baek, Y., Berchuck, S., & Mukherjee, S. (2023). Asymptotics of Bayesian uncertainty estimation in random features regression. *Advances in Neural Information Processing Systems, 36*, 40140-40153.

[7] Rahimi, A., & Recht, B. (2007). Random features for large-scale kernel machines. *Advances in Neural Information Processing Systems, 20*, 1177–1184.

[8] Izmailov, P., Vikram, S., Hoffman, M. D., & Wilson, A. G. G. (2021, July). What are Bayesian neural network posteriors really like?. In *International Conference on Machine Learning* (pp. 4629-4640). PMLR.

[9] Fortuin, V., Garriga-Alonso, A., Wenzel, F., Ratsch, G., Turner, R. E., van der Wilk, M., & Aitchison, L. Bayesian Neural Network Priors Revisited. In *Third Symposium on Advances in Approximate Bayesian Inference*.

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

[11] Minderer, M., Djolonga, J., Romijnders, R., Hubis, F., Zhai, X., Houlsby, N., ... & Lucic, M. (2021). Revisiting the calibration of modern neural networks. *Advances in Neural Information Processing Systems, 34*, 15682-15694.

[12] Catoni, O. (2007). *Pac-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*. stat, 1050, 3.

[13] Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences, 116*(32), 15849-15854.

How to cite this article: Jun Li, Xiao Bai and Jin Zheng (2024). The Cold Posterior Effect in Random Features Models: A Theoretical Explanation. *Bulletin of Computer and Data Sciences*, 5(2), 30-46. DOI: 10.71448/bcds2452-3

Received: 05/05/2024 **Revised:** 29/5/2024 **Accepted:** 20/06/2024 **Publish:** 30/06/2024

Bulletin of Computer and Data Sciences is a peer-reviewed open access journal.