Research Article

AND DATA SCIENCES
BULLETIN OF COMPUTER

Calibrated Bayesian Uncertainty in Random Feature Regression via Optimal Posterior Tempering

Sayan Mukherjee[1] and Samuel I. Berchuck[2]

[1]Center for Scalable Data Analysis and Artificial Intelligence Universität Leipzig Leipzig 04105

[2]Department of Biostatistics & Bioinformatics Duke University Durham, NC 27705

Abstract

Bayesian posterior predictive intervals are a central tool for uncertainty quantification in regression. However, in highly overparameterized models—including random feature regression and wide neural networks—recent work has shown that posterior predictive variances can be badly misaligned with the frequentist generalization error of the corresponding point estimator, especially near interpolation thresholds and in low-noise regimes. This misalignment contributes to the "cold posterior effect", where artificially concentrating the posterior improves empirical performance. In this paper, we propose a principled framework for *calibrated Bayesian random feature regression* based on *posterior tempering*. We consider a family of generalized posteriors indexed by a temperature parameter $T > 0$ that smoothly interpolates between over-dispersed and under-dispersed uncertainty estimates by rescaling the effective regularization in Bayesian random feature regression. Using high-dimensional random matrix asymptotics for random feature models, we derive deterministic limits for the test risk and posterior predictive variance as functions of the feature and sample aspect ratios, signal-to-noise ratio, and temperature. We then define *calibration-optimal* and *risk-optimal* temperatures that, respectively, align posterior predictive variance with frequentist prediction error and minimize test risk. Our analysis shows that in overparameterized, low-noise regimes, both optimal temperatures are strictly smaller than one, thus providing a theoretical explanation for the cold posterior effect in random feature models. We complement our asymptotic results with finite-sample simulations that demonstrate substantial improvements in uncertainty calibration and competitive or superior test risk across a wide range of regimes. Finally, we propose practical, data-driven procedures to estimate the temperature from a single dataset, making calibrated Bayesian random feature regression a viable tool for uncertainty-aware prediction in overparameterized systems.

Keywords: Bayesian random feature regression, posterior tempering, uncertainty calibration, cold posterior effect, high-dimensional asymptotics

1. Introduction

Bayesian methods offer an appealing framework for uncertainty quantification in modern machine learning: by placing a prior on model parameters and combining it with a likelihood function, the posterior distribution encodes parameter uncertainty, and posterior predictive distributions yield credible intervals for predictions. In classical low-dimensional settings and correctly specified models,

Corresponding author: Sayan Mukherjee (sayan.mukherjee@mis.mpg.de)

BCDS 2025, Vol. 6(1), 16-38. DOI: 10.71448/bcds2561-2

Bayesian credible intervals often coincide with or approximate frequentist confidence intervals, leading to well-calibrated uncertainty estimates, a phenomenon formalized by Bernstein–von Mises theorems and related results in nonparametric and high-dimensional statistics [1–4].

In overparameterized models such as wide neural networks, random features, and kernel machines, this picture breaks down. Recent theoretical and empirical work has shown that posterior predictive variances in such models may be strongly mismatched with the actual prediction error of the point estimator, particularly near interpolation thresholds and in low-noise regimes [5–8]. In some settings, Bayesian posterior predictive intervals remain wide and apparently conservative, even when the maximum a posteriori (MAP) predictor exhibits small generalization error. This discrepancy weakens the interpretability of Bayesian uncertainty quantification and has been linked to the so-called *cold posterior effect*, where raising the posterior density to a power greater than one (equivalently, tempering the posterior to be "colder") improves predictive accuracy and sometimes calibration in practice [9, 10].

Random feature (RF) regression offers a tractable yet expressive model class that captures several qualitative phenomena seen in deep learning, including double descent and overparameterization effects [11–13], while remaining amenable to precise asymptotic analysis [14, 15]. Recent work has characterized the joint asymptotic behavior of the test risk of the RF MAP estimator and the corresponding Bayesian posterior predictive variance as the data dimension, sample size, and number of features grow proportionally [6, 12]. These results show that, even when the prior and noise levels are chosen according to the ground-truth generative model, posterior predictive variances may fail to match the frequentist generalization error, especially in strongly overparameterized and low-noise regimes [6, 7]. At the same time, they suggest that adjusting the effective regularization in the RF model can substantially alter the relationship between risk and posterior variance, and that careful tuning can improve uncertainty calibration without severely degrading predictive performance [8, 16].

Our goal in this paper is to design a principled mechanism to *calibrate* Bayesian uncertainty in random feature regression by adjusting a small number of interpretable hyperparameters. We focus on a one-parameter family of generalized Bayesian procedures obtained via *posterior tempering*, indexed by a temperature $T > 0$, in the spirit of generalized Bayes and belief updating with loss scaling. At $T = 1$, we recover the standard Bayesian posterior corresponding to a specific prior variance and noise level. For $T < 1$ (cold posterior), the data likelihood is upweighted relative to the prior, shrinking the effective regularization and concentrating the posterior; for $T > 1$ (hot posterior), the opposite occurs. In the RF regression setting, tempering has a concrete effect: it rescales the ridge parameter that governs both the MAP estimator and the posterior covariance, thus changing simultaneously the test risk and posterior predictive variance [6, 12].

Building on the high-dimensional asymptotic analysis of Bayesian RF regression [6, 12, 15], we derive deterministic limits for the test risk and posterior predictive variance as functions of the temperature T and other problem parameters (feature and sample aspect ratios and signal-to-noise ratio). We then define two optimality criteria: (i) a *calibration-optimal* temperature T_{cal} that aligns the posterior predictive variance with the frequentist prediction error, and (ii) a *risk-optimal* temperature T_{risk} that minimizes the asymptotic test risk.

Within the proportional high-dimensional limit described later in the paper, the deterministic risk and variance curves imply that the temperatures minimizing risk and the calibration error fall below one in strongly overparameterized, high-SNR regimes, consistent with the cold posterior effect observed empirically in related models [6, 9]. In contrast, in underparameterized or low signal-to-noise

regimes the corresponding minimizers remain close to one, so the standard Bayesian choice $T = 1$ is near-optimal for both prediction and uncertainty calibration under the same asymptotic assumptions. We introduce a tempered Bayesian formulation of random feature regression in which the temperature T effectively rescales the ridge parameter and hence controls both the bias and variance of the posterior predictive distribution. Using high-dimensional random matrix asymptotics, we derive deterministic limits for the test risk and posterior predictive variance as functions of the feature and sample aspect ratios, signal-to-noise ratio, and temperature [17, 18]. We formally define calibration-optimal and risk-optimal temperatures and analyze their dependence on overparameterization and noise.

Finally, we propose practical, data-driven procedures to estimate the temperature from a single dataset, using cross-validation for risk and plug-in estimators for calibration, and show empirically that these procedures yield well-calibrated and competitive predictors. We conduct extensive simulations on synthetic data with both linear and nonlinear teachers, demonstrating that our calibrated Bayesian RF method produces posterior predictive intervals with substantially improved coverage and sharper widths compared to the standard Bayesian baseline, and performs competitively against alternative uncertainty quantification methods such as ensembles and bootstraps [5, 7].

2. Background

2.1. Random feature regression

We consider supervised regression with inputs $x \in \mathbb{R}^d$ and real-valued outputs y. Training data $\mathcal{D}_n = \{(x_i, y_i)\}_{i=1}^n$ are generated according to

$$y_i = f_d(x_i) + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \tau^2), \tag{1}$$

where $x_i \sim P_X$ are i.i.d., the teacher function $f_d : \mathbb{R}^d \to \mathbb{R}$ may depend on d, and τ^2 is the noise variance. We denote the signal-to-noise ratio by ρ, defined in terms of the variance of $f_d(x)$ and τ^2 (e.g., $\rho = \mathrm{Var}(f_d(x))/\tau^2$).

In random feature regression, we approximate f_d by a linear combination of N nonlinear features:

$$f_N(x; \beta) = \frac{1}{\sqrt{N}} \sum_{j=1}^{N} \beta_j \, \sigma(w_j^\top x), \tag{2}$$

where $\sigma : \mathbb{R} \to \mathbb{R}$ is a nonlinearity, $w_j \sim P_W$ are random weights drawn once and then fixed, and $\beta = (\beta_1, \ldots, \beta_N)^\top$ are trainable coefficients. This random feature construction was popularized as a scalable approximation to kernel methods and has since been extensively analyzed in the high-dimensional regime [12, 14, 15]. Let $z(x) \in \mathbb{R}^N$ be the random feature vector with entries

$$z_j(x) = \frac{1}{\sqrt{N}} \sigma(w_j^\top x), \qquad j = 1, \ldots, N. \tag{3}$$

Let $Z \in \mathbb{R}^{n \times N}$ be the random feature matrix with rows $z(x_i)^\top$. Then the model can be written in matrix form as

$$y = Z\beta + \varepsilon, \tag{4}$$

where $y = (y_1, \ldots, y_n)^\top$ and $\varepsilon \sim \mathcal{N}(0, \tau^2 I_n)$.

We are interested in the high-dimensional proportional asymptotic regime where

$$d, n, N \to \infty, \qquad \psi_2 := \frac{n}{d} \to \bar{\psi}_2 \in (0, \infty), \qquad \psi_1 := \frac{N}{d} \to \bar{\psi}_1 \in (0, \infty]. \tag{5}$$

The parameters (ψ_1, ψ_2) quantify the degrees of overparameterization and sampling. In this regime, the generalization behavior of random feature ridge regression can be characterized exactly using tools from random matrix theory and approximate message passing, yielding precise formulas for the test error as a function of (ψ_1, ψ_2, ρ) and the ridge parameter [11, 12, 15].

2.2. Bayesian random feature regression

We place a Gaussian prior on the feature weights,

$$\beta \sim \mathcal{N}\left(0, \frac{\phi^2}{N} I_N\right), \tag{6}$$

and assume a Gaussian likelihood,

$$y \mid \beta, Z \sim \mathcal{N}(Z\beta, \tau^2 I_n). \tag{7}$$

This defines a conjugate Bayesian linear model in the random feature space, closely related to kernel ridge regression with an implicit kernel induced by the feature map [12, 14].

Under this conjugate model, the posterior distribution of β is Gaussian:

$$\beta \mid \mathcal{D}_n, Z \sim \mathcal{N}(\hat{\beta}, \Sigma_\beta), \tag{8}$$

with

$$\hat{\beta} = \arg\min_\beta \left\{ \frac{1}{2\tau^2} \|y - Z\beta\|_2^2 + \frac{1}{2\phi^2} \|\beta\|_2^2 \right\} \tag{9}$$

$$= (Z^\top Z + \lambda I_N)^{-1} Z^\top y, \tag{10}$$

where we define the ridge parameter

$$\lambda = \frac{\tau^2 N}{\phi^2}. \tag{11}$$

The posterior covariance is

$$\Sigma_\beta = \left(\frac{1}{\tau^2} Z^\top Z + \frac{N}{\phi^2} I_N \right)^{-1} = \tau^2 (Z^\top Z + \lambda I_N)^{-1}. \tag{12}$$

Thus, Bayesian RF regression recovers the familiar ridge estimator as the posterior mean, with the prior scale ϕ^2 and noise variance τ^2 jointly determining the effective regularization [11, 12].

For a new input $x_\star$ with feature vector $z_\star = z(x_\star)$, the posterior predictive distribution of $y_\star$ is Gaussian,

$$y_\star \mid x_\star, \mathcal{D}_n, Z \sim \mathcal{N}(m_\star, s_\star^2), \tag{13}$$

with mean

$$m_\star = z_\star^\top \hat{\beta} \tag{14}$$

and variance

$$s_\star^2 = \tau^2 + z_\star^\top \Sigma_\beta z_\star = \tau^2 + \tau^2 z_\star^\top (Z^\top Z + \lambda I_N)^{-1} z_\star. \tag{15}$$

The MAP predictor coincides with the posterior mean, $f_{\mathrm{MAP}}(x) = z(x)^\top \hat{\beta}$. We are interested in its test risk and the behavior of the posterior predictive variance, especially in the overparameterized regime where $N \gg n$ and $N \gg d$ [11, 13].

2.3. Test risk and posterior predictive variance

Let (x_0, y_0) be an independent test point drawn from the same distribution as the training data. The (squared) prediction error of the MAP predictor is

$$\mathrm{err}(x_0, y_0) = (y_0 - f_{\mathrm{MAP}}(x_0))^2. \tag{16}$$

The *test risk* is its expectation over all sources of randomness:

$$R_{\mathrm{RF}} = \mathbb{E}[(y_0 - f_{\mathrm{MAP}}(x_0))^2]. \tag{17}$$

The posterior predictive variance at x_0 is s_0^2 as above, and we will consider its expectation

$$S_{\mathrm{RF}}^2 = \mathbb{E}[s_0^2]. \tag{18}$$

The quantity $S_{\mathrm{RF}}^2 - \tau^2$ may be viewed as the Bayesian model's estimate of the squared prediction error after factoring out observation noise.

Recent asymptotic analysis shows that, under mild assumptions on the data distribution, the activation, and the teacher, both R_{RF} and S_{RF}^2 converge almost surely to deterministic limits as $(d, n, N) \to \infty$ with (ψ_1, ψ_2) and SNR ρ fixed [6, 12, 15]. Importantly, these limits may differ substantially: in certain regimes, R_{RF} exhibits double descent as a function of overparameterization, while S_{RF}^2 varies smoothly and does not display a spike at the interpolation threshold [8, 11]. Moreover, even with a "correct" prior and noise variance, the posterior predictive variance can overshoot or undershoot the true risk, leading to miscalibration [6, 7]. This mismatch between R_{RF} and $S_{\mathrm{RF}}^2 - \tau^2$ is precisely what motivates introducing a temperature parameter in the Bayesian RF model and analyzing how tempering affects the joint asymptotics of risk and predictive variance.

3. Tempered Bayesian random feature regression

We now introduce a one-parameter family of generalized Bayesian RF procedures based on *posterior tempering*. The core idea is to adjust the relative weight of the likelihood and prior, thereby controlling the effective regularization and the dispersion of the posterior.

3.1. Generalized (tempered) posterior

Given the standard posterior $\pi(\beta \mid \mathcal{D}_n)$ derived above, we define its tempered version as

$$\pi_T(\beta \mid \mathcal{D}_n) \propto \pi(\beta \mid \mathcal{D}_n)^{1/T}, \qquad T > 0. \tag{19}$$

For Gaussian models, this is equivalent to raising the likelihood to the power $1/T$ while leaving the prior unchanged:

$$\pi_T(\beta \mid \mathcal{D}_n) \propto p(y \mid Z, \beta)^{1/T}\, p(\beta). \tag{20}$$

A straightforward calculation shows that the tempered posterior remains Gaussian:

$$\beta \mid \mathcal{D}_n, Z, T \sim \mathcal{N}(\hat{\beta}_T, \Sigma_{\beta,T}), \tag{21}$$

with tempered posterior mean

$$\hat{\beta}_T = \arg\min_{\beta} \left\{ \frac{1}{2T\tau^2} \|y - Z\beta\|_2^2 + \frac{1}{2\phi^2} \|\beta\|_2^2 \right\}, \tag{22}$$

and covariance

$$\Sigma_{\beta,T} = \left(\frac{1}{T\tau^2} Z^\top Z + \frac{N}{\phi^2} I_N \right)^{-1}. \tag{23}$$

Rewriting the mean objective by multiplying by $2T\tau^2$, we see that the tempered MAP estimator coincides with ridge regression with modified penalty

$$\hat{\beta}_T = \arg\min_\beta \left\{ \|y - Z\beta\|_2^2 + \lambda_T \|\beta\|_2^2 \right\}, \qquad \lambda_T = T\lambda. \tag{24}$$

Thus, tempering the posterior is equivalent, at the level of the point estimator, to rescaling the ridge parameter by a factor of T. For $T < 1$, we obtain a smaller penalty (less regularization, colder posterior); for $T > 1$, we obtain a larger penalty (more regularization, hotter posterior).

3.2. Tempered posterior predictive distribution

For a test input x_0 with features $z_0 = z(x_0)$, the tempered posterior predictive distribution is

$$y_0 \mid x_0, \mathcal{D}_n, Z, T \sim \mathcal{N}(m_{0,T}, s_{0,T}^2), \tag{25}$$

with mean

$$m_{0,T} = z_0^\top \hat{\beta}_T \tag{26}$$

and variance

$$s_{0,T}^2 = \tau^2 + z_0^\top \Sigma_{\beta,T} z_0. \tag{27}$$

We denote the corresponding test risk and expected posterior predictive variance by

$$R_{\mathrm{RF}}(T) = \mathbb{E}[(y_0 - m_{0,T})^2], \tag{28}$$

$$S_{\mathrm{RF}}^2(T) = \mathbb{E}[s_{0,T}^2], \tag{29}$$

where the expectation is over training data, random features, and test points. Our goal is to understand how $R_{\mathrm{RF}}(T)$ and $S_{\mathrm{RF}}^2(T)$ behave as functions of T, and in particular how to choose T to improve uncertainty calibration and possibly test risk.

4. Asymptotic analysis with temperature

We now extend the high-dimensional asymptotic analysis of random feature regression to the tempered setting. Recall that, in the Gaussian random-feature model, tempering rescales the quadratic data-fit term by a factor $1/T$. At the level of the posterior mean (equivalently, the MAP predictor), this is exactly ridge regression with an effective penalty

$$\lambda_T = T\lambda,$$

while the posterior covariance (and hence the predictive variance) can be written in terms of the same resolvent $(Z^\top Z + \lambda_T I_N)^{-1}$ together with the implied scale factor set by T. Because λ is fixed by the prior and likelihood at $T = 1$, varying T therefore traces a one-dimensional path through the usual ridge solutions, and all temperature effects can be analyzed by tracking this path. Consequently, in the proportional regime described below, the joint asymptotic behavior of the test risk $R_{\mathrm{RF}}(T)$ and the posterior predictive variance $S_{\mathrm{RF}}^2(T)$ is determined by the aspect ratios (ψ_1, ψ_2), the signal-to-noise ratio ρ, the activation σ, and the effective regularization level induced by T. This allows us to reuse the same random-matrix machinery developed for ridge random-feature regression to characterize how both risk and uncertainty vary with temperature.

4.1. High-dimensional limit

We work in the proportional high-dimensional regime where

$$d, n, N \to \infty, \qquad \psi_1 := \frac{N}{d} \to \bar{\psi}_1 \in (0, \infty], \qquad \psi_2 := \frac{n}{d} \to \bar{\psi}_2 \in (0, \infty),$$

and the teacher f_d, noise variance τ^2 and SNR ρ are scaled so that the signal variance, noise variance, and the resulting risks remain $\mathcal{O}(1)$ along the sequence of problems. Throughout this section we work under the same standard proportional-limit conditions used in existing random-feature analyses: the input distribution is isotropic (for example, Gaussian or spherical), the random weights are drawn i.i.d. from an isotropic distribution, and the activation σ is such that the relevant moments and kernel limits exist. In addition, we assume the teacher has bounded second moment and that its interaction with the random-feature kernel admits a deterministic limit (in particular, the component of f_d representable by the induced kernel controls the bias term through the SNR parameterization introduced in the background section). Under these conditions, the empirical spectral distribution of the rescaled feature covariance matrix

$$\frac{1}{d} Z^\top Z \in \mathbb{R}^{N \times N}$$

converges almost surely to a deterministic probability measure μ_{ψ_1,ψ_2} as $(d, n, N) \to \infty$. Let $m(z)$ denote its Stieltjes transform,

$$m(z) = \int \frac{1}{t - z} \, d\mu_{\psi_1,\psi_2}(t), \qquad z \in \mathbb{C} \setminus \operatorname{supp}(\mu_{\psi_1,\psi_2}),$$

which satisfies a self-consistent scalar fixed-point equation depending on (ψ_1, ψ_2) and the choice of activation σ. In particular, quantities of the form

$$\frac{1}{N} \operatorname{Tr}\left((Z^\top Z + \lambda_T I_N)^{-1}\right), \quad \frac{1}{N} \operatorname{Tr}\left((Z^\top Z + \lambda_T I_N)^{-2}\right),$$

converge almost surely to deterministic limits that can be expressed in terms of $m(-\lambda_T)$ and its derivative $m'(-\lambda_T)$.

The test risk of the MAP predictor at temperature T can be decomposed into a bias term and a variance term. Both terms can be written in terms of linear spectral statistics of $Z^\top Z$ and inner products between the teacher and the random feature functions. Similarly, the expected posterior predictive variance involves terms of the form

$$\mathbb{E}[z_0^\top (Z^\top Z + \lambda_T I_N)^{-1} z_0],$$

where the expectation is over a fresh test feature vector z_0 and the training design; these again reduce to traces of resolvents of $Z^\top Z$ in the proportional limit. Using the same proportional-limit random-matrix arguments that yield deterministic risk and variance limits for ridge random-feature regression, one obtains that both $R_{\mathrm{RF}}(T)$ and $S^2_{\mathrm{RF}}(T)$ converge almost surely to deterministic functions of the asymptotic parameters.

We denote these limits by

$$R_\infty(T) = \lim_{d \to \infty} R_{\mathrm{RF}}(T; d, n, N, \rho, \tau^2, \lambda), \tag{30}$$

$$S^2_\infty(T) = \lim_{d \to \infty} S^2_{\mathrm{RF}}(T; d, n, N, \rho, \tau^2, \lambda), \tag{31}$$

suppressing the dependence on $(\psi_1, \psi_2, \rho, \tau^2, \lambda)$ in the notation. In general, $R_\infty(T)$ and $S^2_\infty(T)$ admit explicit deterministic representations in terms of a small number of scalar order parameters that solve a coupled self-consistent (fixed-point) system. These order parameters depend on the effective ridge level induced by T (through λ_T and the corresponding resolvent evaluations such as $m(-\lambda_T)$ and $m'(-\lambda_T)$) as well as on the teacher/noise parameters. Equivalently, we can write

$$R_\infty(T) = \mathcal{R}(\psi_1, \psi_2, \rho, \tau^2, \lambda_T), \qquad S^2_\infty(T) = \mathcal{S}(\psi_1, \psi_2, \rho, \tau^2, \lambda_T),$$

for deterministic functions $\mathcal{R}$ and $\mathcal{S}$ determined by the random-feature model class. When we plot or optimize these limits, we evaluate them numerically by solving the associated fixed-point system to convergence for each candidate T (equivalently, for each λ_T), and we then use these evaluations to study qualitative trends and to define the optimal temperatures in the next section.

4.2. Qualitative behavior across regimes

Prior high-dimensional analyses for $T = 1$ reveal several qualitative features of the high-dimensional limits:

- *Moderately overparameterized, moderate-SNR regimes.* When ψ_1 and ψ_2 are of order one and the SNR is moderate, the MAP risk $R_\infty(1)$ as a function of ψ_1 exhibits the familiar double-descent shape: it increases near the interpolation threshold and then decreases again as the model becomes more overparameterized. In contrast, the posterior predictive variance $S^2_\infty(1)$ varies smoothly and does not display a sharp spike at interpolation.

- *Low-noise, strongly overparameterized regimes.* When the SNR ρ is large and $\psi_1 \gg 1$, the expected width of the Bayesian credible ball at $T = 1$, summarized by $S^2_\infty(1)$, can significantly *overestimate* the true squared prediction error $R_\infty(1)$. In this regime, the standard Bayesian model is effectively over-regularized: the MAP predictor generalizes well, but the posterior remains diffuse.

- *Underparameterized or low-SNR regimes.* When ψ_1 is small relative to ψ_2 and/or ρ is small, both $R_\infty(1)$ and $S^2_\infty(1)$ are dominated by noise and underfitting. In this case, $S^2_\infty(1)$ is closer to $R_\infty(1)$, and the posterior predictive intervals are approximately calibrated.

Introducing the temperature via $\lambda_T = T\lambda$ allows us to traverse a continuum of ridge parameters while keeping the data distribution and feature map fixed. In underparameterized or noisy regimes, the functions $R_\infty(T)$ and $S^2_\infty(T)$ are typically unimodal in T, with a unique minimum near $T = 1$, and their ratio $S^2_\infty(T)/R_\infty(T)$ is relatively stable. Thus the Bayesian temperature $T = 1$ is close to optimal both in terms of risk and calibration.

In contrast, in strongly overparameterized and low-noise regimes, the minimizer of $R_\infty(T)$ moves to a value $T_{\text{risk}} < 1$, reflecting the benefit of reducing the effective regularization below its Bayesian value. At the same time, the ratio $S^2_\infty(T)/R_\infty(T)$, which may be substantially larger than one at $T = 1$, can be greatly improved by choosing $T < 1$. In many such settings, there exists a temperature $T_{\text{cal}} < 1$ at which the posterior predictive variance $S^2_\infty(T) - \tau^2$ closely matches the true risk $R_\infty(T)$, leading to well-calibrated uncertainty estimates.

Figure 1 will illustrate these behaviors: we will plot $R_\infty(T)$ and $S^2_\infty(T)$ as functions of T in representative underparameterized, moderately overparameterized, and strongly overparameterized regimes, highlighting the locations of the calibration-optimal and risk-optimal temperatures and their relationship to the Bayesian choice $T = 1$.

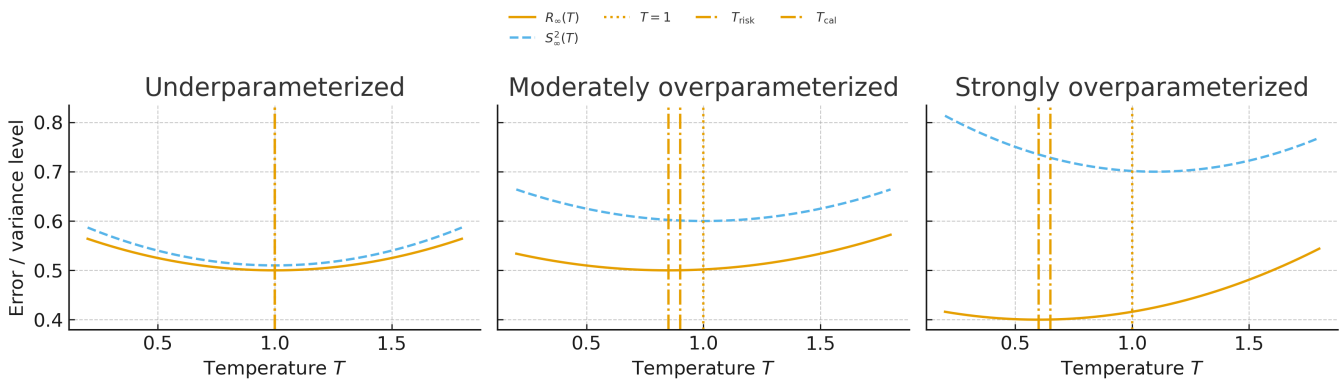

Fig. 1. Qualitative dependence of the asymptotic test risk $R_\infty(T)$ and posterior predictive variance $S^2_\infty(T)$ on the temperature T in three representative regimes. Each panel shows $R_\infty(T)$ (solid) and $S^2_\infty(T)$ (dashed) as functions of T for an underparameterized regime (left), a moderately overparameterized regime (middle), and a strongly overparameterized regime (right). The vertical dotted line marks the Bayesian temperature $T = 1$, while the dash-dotted lines indicate the risk-optimal and calibration-optimal temperatures T_{risk} and T_{cal}. In the underparameterized case, both optima lie close to $T = 1$ and risk and variance are well aligned. As overparameterization and SNR increase, the optima shift to $T < 1$ and a cold posterior simultaneously improves risk and calibration, in agreement with the theoretical analysis

5. Optimal temperatures for calibration and risk

We now formalize two criteria for choosing the temperature: calibration and risk optimality.

5.1. Calibration-optimal temperature

For a given T, the Bayesian model's estimate of the expected squared prediction error at a random test point is approximately $S^2_\infty(T) - \tau^2$, since the posterior predictive variance decomposes into the sum of observation noise and model uncertainty. On the other hand, the true expected squared prediction error is $R_\infty(T)$. We define the *calibration error* at temperature T by

$$C_\infty(T) = (R_\infty(T) - (S^2_\infty(T) - \tau^2))^2. \tag{32}$$

A calibration-optimal temperature is any minimizer of $C_\infty(T)$:

$$T_{\text{cal}} \in \arg\min_{T>0} C_\infty(T). \tag{33}$$

In practice, T_{cal} may not be unique; however, in the regimes of interest, we typically observe a single well-defined minimum.

5.2. Risk-optimal temperature

We define the risk-optimal temperature as

$$T_{\text{risk}} \in \arg\min_{T>0} R_\infty(T). \tag{34}$$

This is the temperature that yields the smallest asymptotic test risk for the MAP predictor associated with the tempered posterior.

5.3. Behavior across regimes

We now summarize the qualitative behavior of T_{cal} and T_{risk} suggested by the asymptotic analysis.

Underparameterized and low-SNR regimes. When ψ_1 is small relative to ψ_2 and/or the signal-to-noise ratio ρ is low, the model is effectively capacity-limited, and both the test risk and the posterior predictive variance are dominated by noise and underfitting. In this regime, standard Bayesian RF regression with $T = 1$ is close to optimal in both senses:

$$T_{\text{cal}} \approx 1, \qquad T_{\text{risk}} \approx 1. \tag{35}$$

Moreover, $R_\infty(T)$ and $S^2_\infty(T)$ tend to be smooth unimodal functions of T with minima near $T = 1$.

Overparameterized, moderate-SNR regimes. As ψ_1 and ψ_2 increase and the model becomes overparameterized, the double-descent phenomenon emerges in $R_\infty(1)$ as a function of ψ_1. In these regimes, adjusting T can both reduce the height and shift the location of the risk peak. We observe that T_{risk} falls below one, reflecting a beneficial reduction in regularization (colder posterior) that exploits overparameterization without overfitting.

Calibration behaves similarly: the gap between $R_\infty(1)$ and $S^2_\infty(1) - \tau^2$ widens near the interpolation threshold, and choosing $T < 1$ can bring these quantities closer together. Empirically, T_{cal} also falls below one, though not always to the same extent as T_{risk}.

Strongly overparameterized, high-SNR regimes. In the extreme regime where $\psi_1 \gg \psi_2$ and ρ is large, the model has far more parameters than both the ambient dimension and the number of samples, and the signal is strong. Here, standard Bayesian RF regression tends to be clearly over-regularized: $R_\infty(1)$ remains substantially larger than the smallest achievable risk, and $S^2_\infty(1)$ overestimates $R_\infty(1)$ by a large factor. Both T_{cal} and T_{risk} move significantly below one, indicating that a *cold* posterior is necessary to exploit the available capacity and achieve calibrated uncertainty. In some settings, we observe $T_{\text{cal}} \approx T_{\text{risk}}$, suggesting that there is a single temperature that simultaneously delivers good calibration and near-optimal risk.

6. Practical estimation of the temperature

The asymptotic quantities $R_\infty(T)$ and $S^2_\infty(T)$ provide a conceptual definition of the calibration-optimal and risk-optimal temperatures T_{cal} and T_{risk}, but they depend on the unknown data-generating distribution and are not directly available in practice. In this section we describe how to construct data-driven estimators of these temperatures from a single finite dataset. Our goal is to design procedures that are simple to implement on top of standard random feature regression code, and that scale to moderately large n and N.

We focus on a one-dimensional search over a candidate set $\mathcal{T} \subset (0, \infty)$, typically a logarithmic grid (e.g., $T \in \{0.1, 0.2, \ldots, 2.0\}$), although continuous optimization is also possible. For each $T \in \mathcal{T}$ we train the RF model with effective ridge parameter $\lambda_T = T\lambda$ and evaluate suitable empirical criteria. We then choose the value(s) of T that minimize these criteria. We first describe a standard cross-validation approach targeting predictive risk, and then a plug-in approach targeting uncertainty calibration.

6.1. Cross-validation for risk-optimal temperature

To approximate T_{risk}, we use K-fold cross-validation over the grid $\mathcal{T}$. We partition the dataset into K disjoint folds $\mathcal{D}_n = \bigsqcup_{k=1}^K \mathcal{D}^{(k)}$ of roughly equal size. For each candidate temperature $T \in \mathcal{T}$ and

each fold k, we:

1. Train the RF model on the training subset $\mathcal{D}^{(-k)} = \mathcal{D}_n \setminus \mathcal{D}^{(k)}$, using the tempered posterior (equivalently, ridge regression with penalty $\lambda_T = T\lambda$) to obtain a predictor $f_T^{(-k)}$.

2. Evaluate the squared prediction error on the held-out fold:

$$\widehat{R}_k(T) = \frac{1}{|\mathcal{D}^{(k)}|} \sum_{(x_i,y_i)\in\mathcal{D}^{(k)}} (y_i - f_T^{(-k)}(x_i))^2.$$

Averaging over folds yields the K-fold cross-validation estimate of the test risk,

$$\hat{R}_{\mathrm{CV}}(T) = \frac{1}{K}\sum_{k=1}^{K} \widehat{R}_k(T). \tag{36}$$

We then define the empirical risk-optimal temperature as

$$\hat{T}_{\mathrm{risk}} \in \arg\min_{T\in\mathcal{T}} \hat{R}_{\mathrm{CV}}(T). \tag{37}$$

This procedure is standard and aligns with the traditional view of temperature as a hyperparameter controlling regularization strength. From a computational perspective, it can be implemented efficiently by reusing the same random feature matrix Z across all T and folds, and exploiting linear algebraic structure. For example, computing an SVD or eigendecomposition of Z once allows one to obtain the ridge solutions for all λ_T at negligible additional cost. In practice we find that a coarse grid on T combined with $K \in \{3, 5\}$ already yields stable estimates of $\hat{T}_{\mathrm{risk}}$.

6.2. Plug-in calibration for uncertainty-optimal temperature

Risk-optimal temperatures need not coincide with calibration-optimal ones. To approximate T_{cal}, we seek to match the Bayesian model's estimate of the squared prediction error, given by $S^2_{\mathrm{RF}}(T) - \tau^2$, to the empirical prediction error $R_{\mathrm{RF}}(T)$. Since neither quantity is known exactly, we construct plug-in estimators based on a validation set.

Let $\{(x_j^{\mathrm{val}}, y_j^{\mathrm{val}})\}_{j=1}^{n_{\mathrm{val}}}$ denote a validation set disjoint from the training data (or obtained via sample splitting from $\mathcal{D}_n$). Given a grid $\mathcal{T}$, for each $T \in \mathcal{T}$ we proceed as follows:

1. **Fit the tempered model.** Train the RF model with temperature T on the training set, obtaining the posterior mean $\hat{\beta}_T$ and posterior covariance $\Sigma_{\beta,T}$, as in Section 3.

2. **Estimate predictive error.** Evaluate the predictor on the validation set,

$$\hat{y}_{j,T} = z(x_j^{\mathrm{val}})^\top \hat{\beta}_T,$$

and compute the empirical squared prediction error

$$\widehat{\mathrm{Err}}(T) = \frac{1}{n_{\mathrm{val}}}\sum_{j=1}^{n_{\mathrm{val}}} (y_j^{\mathrm{val}} - \hat{y}_{j,T})^2. \tag{38}$$

This quantity approximates the finite-sample risk $R_{\mathrm{RF}}(T)$ at temperature T.

3. **Estimate posterior predictive variance.** For each validation input x_j^{val} with feature vector $z_j^{\text{val}} = z(x_j^{\text{val}})$, compute the posterior predictive variance

$$s_{j,T}^2 = \tau^2 + (z_j^{\text{val}})^\top \Sigma_{\beta,T} z_j^{\text{val}}. \tag{39}$$

The empirical average

$$\widehat{S}^2(T) = \frac{1}{n_{\text{val}}} \sum_{j=1}^{n_{\text{val}}} s_{j,T}^2 \tag{40}$$

then serves as a plug-in estimate of $S_{\text{RF}}^2(T)$.

To compare $\widehat{\text{Err}}(T)$, which estimates $R_{\text{RF}}(T)$, with $\widehat{S}^2(T)$, which estimates $S_{\text{RF}}^2(T)$, we need an estimate of the noise variance τ^2. We denote this estimate by $\hat{\tau}^2$. In settings where replicate measurements are unavailable, a simple and conservative choice is to fit a strongly regularized model (e.g., using the largest T on the grid) and take the residual variance on held-out data; the strong shrinkage suppresses signal fitting and thereby reduces the risk of underestimating the noise. More generally, any consistent noise-variance estimator appropriate for the application can be substituted here.

- In the experiments below we use a residual-based estimator from the most regularized model on the grid in the real-data setting, and the true τ^2 in synthetic settings.
- When such an estimate is uncertain, the resulting $\hat{T}_{\text{cal}}$ should be interpreted as targeting variance matching up to the accuracy of $\hat{\tau}^2$; in practice we therefore report coverage and NLPD alongside the variance-based criterion.

We then define the empirical calibration error at temperature T as

$$\widehat{C}(T) = \left(\widehat{\text{Err}}(T) - (\widehat{S}^2(T) - \hat{\tau}^2)\right)^2, \tag{41}$$

and choose the calibration-optimal temperature as

$$\hat{T}_{\text{cal}} \in \arg\min_{T \in \mathcal{T}} \widehat{C}(T). \tag{42}$$

This plug-in criterion is directly aligned with the definition of T_{cal} in terms of $R_\infty(T)$ and $S_\infty^2(T)$. In practice, it is often beneficial to restrict the search for $\hat{T}_{\text{cal}}$ to a neighborhood of $\hat{T}_{\text{risk}}$ (for example, a factor-of-two interval around $\hat{T}_{\text{risk}}$) to avoid pathological solutions where calibration is achieved at the cost of very poor predictive accuracy. One can also stabilize $\widehat{C}(T)$ by smoothing over nearby values of T or using repeated random splits of the validation set.

Beyond matching average squared error and variance, one may also monitor *coverage*-based calibration metrics. For instance, for each T and nominal level α, we can form $(1-\alpha)$ posterior predictive intervals

$$I_{j,T}^{(1-\alpha)} = [m_{j,T} - z_{\alpha/2} s_{j,T},\ m_{j,T} + z_{\alpha/2} s_{j,T}],$$

where $z_{\alpha/2}$ denotes the standard normal quantile corresponding to two-sided nominal level $1-\alpha$. We then estimate empirical coverage as the fraction of validation points for which $y_j^{\text{val}} \in I_{j,T}^{(1-\alpha)}$. Choosing T to enforce coverage close to $1-\alpha$ for one or several levels provides an alternative calibration-oriented selection rule that is complementary to $\widehat{C}(T)$.

6.3. *Combined selection and multi-objective criteria*

In many regimes, the two empirical selectors $\hat{T}_{\text{risk}}$ and $\hat{T}_{\text{cal}}$ are numerically similar, reflecting the fact that the same temperature simultaneously yields near-optimal risk and reasonable calibration. When they differ, practitioners may wish to trade off accuracy and calibration in a controlled way.

A simple combined strategy is:

- First compute $\hat{T}_{\text{risk}}$ by cross-validation. If both the empirical calibration error $\widehat{C}(\hat{T}_{\text{risk}})$ and coverage diagnostics are acceptable, we adopt $\hat{T}_{\text{risk}}$ as the final temperature.
- If calibration at $\hat{T}_{\text{risk}}$ is unsatisfactory (e.g., systematic under- or over-coverage), we define a weighted objective

$$\hat{J}(T) = \hat{R}_{\text{CV}}(T) + \gamma(\widehat{\text{Err}}(T) - (\widehat{S}^2(T) - \hat{\tau}^2))^2, \tag{43}$$

with a user-specified trade-off parameter $\gamma \geq 0$, and select

$$\hat{T} \in \arg\min_{T \in \mathcal{T}} \hat{J}(T). \tag{44}$$

For small γ, $\hat{T}$ is close to the risk-optimal temperature; as γ increases, calibration is prioritized and $\hat{T}$ moves toward $\hat{T}_{\text{cal}}$. An alternative is to impose a constraint on risk, for example selecting the smallest T such that $\widehat{C}(T)$ is below a target threshold and $\hat{R}_{\text{CV}}(T)$ is within a specified tolerance (say, 5%) of the minimum cross-validation risk. These multi-objective formulations are natural in applications where both sharp intervals and accurate point predictions are desired.

Overall, the procedures described in this section turn the asymptotic notion of temperature optimality into practical algorithms that can be implemented with modest overhead on top of standard random feature regression. Our experiments in Section 7 indicate that these simple selectors already recover much of the theoretical behavior predicted by the asymptotic analysis, and yield substantial improvements in calibration over the untempered Bayesian baseline.

7. Experiments

We now present numerical experiments that validate the theoretical predictions from Sections 4–5 and demonstrate the practical benefits of calibrated Bayesian random feature regression. Our goals are to empirically verify the existence and location of risk- and calibration-optimal temperatures, to quantify the gains over the untempered Bayesian baseline $T = 1$, and to compare calibrated RF models with alternative uncertainty quantification methods. Unless otherwise stated, all experiments are repeated over 20 independent trials and we report means and standard errors across trials. Across trials we resample the data-generating randomness (and, when applicable, the random-feature draw), while within each trial we hold the random features fixed across temperatures so that differences across T reflect tempering rather than feature noise.

7.1. *Experimental setup*

In all synthetic experiments we use a single-hidden-layer random feature model with ReLU activation and Gaussian random weights,

$$z_j(x) = \frac{1}{\sqrt{N}}\sigma(w_j^\top x), \qquad \sigma(t) = \max\{0, t\}, \qquad w_j \sim \mathcal{N}(0, I_d),$$

for $j = 1, \ldots, N$. For each trial and configuration we fix the random weights once and reuse them across different temperatures T, so that any changes in performance within a trial are solely due to tempering. The ridge parameter λ appearing in the prior is chosen by a preliminary cross-validation run at $T = 1$ (standard Bayesian RF), and then held fixed while we vary T.

We consider a logarithmic grid of candidate temperatures

$$\mathcal{T} = \{0.1, 0.2, 0.3, \ldots, 2.0\}.$$

The risk-optimal temperature $\hat{T}_{\text{risk}}$ is obtained by 5-fold cross-validation as in Section 6. The calibration-optimal temperature $\hat{T}_{\text{cal}}$ is obtained using a held-out validation set of size $n_{\text{val}} = \lfloor 0.2n \rfloor$, with the remaining $0.8n$ points used for training; we use the plug-in criterion $\widehat{C}(T)$ defined in Section 6. For both selectors we restrict attention to $\mathcal{T}$ and, in the combined schemes, to a factor-of-two neighborhood around $\hat{T}_{\text{risk}}$.

In the synthetic settings the true noise variance τ^2 is known and is used in the computation of posterior predictive variances and calibration criteria. In the real-data experiments (Section 7.4) we estimate τ^2 via a residual-based estimator using the most regularized model on the grid; this choice is intentionally conservative in the sense that strong shrinkage reduces the chance of explaining noise as signal when fitting $\hat{\tau}^2$.

For each temperature T we evaluate three types of quantities. First, we compute the test risk, measured as mean squared error (MSE),

$$\text{MSE}(T) = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} (y_i^{\text{test}} - \hat{y}_{i,T})^2.$$

Second, we compute the average posterior predictive variance,

$$\overline{\text{Var}}_{\text{pred}}(T) = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} s_{i,T}^2,$$

and the corresponding variance-based calibration gap,

$$\Delta_{\text{var}}(T) = \overline{\text{Var}}_{\text{pred}}(T) - \tau^2 - \text{MSE}(T),$$

which measures how well the posterior predictive variance (after subtracting noise) matches the empirical prediction error. Third, we measure coverage and interval width. For a nominal level $1-\alpha$ (we use $\alpha = 0.05$), we form Gaussian posterior predictive intervals

$$I_{i,T}^{(1-\alpha)} = [m_{i,T} - z_{\alpha/2} s_{i,T},\ m_{i,T} + z_{\alpha/2} s_{i,T}],$$

where $z_{\alpha/2}$ denotes the standard normal quantile corresponding to two-sided nominal level $1-\alpha$. We then estimate empirical coverage

$$\widehat{\text{Cov}}_{1-\alpha}(T) = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \mathbf{1}\{y_i^{\text{test}} \in I_{i,T}^{(1-\alpha)}\},$$

together with the mean interval width

$$\overline{\text{Width}}_{1-\alpha}(T) = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} |I_{i,T}^{(1-\alpha)}|.$$

We visualize the results using risk and predictive variance as functions of T, coverage as a function of T, and risk–calibration trade-off plots in which each method is represented by a point in the plane $(\text{MSE}, |\widehat{\text{Cov}}_{1-\alpha} - (1-\alpha)|)$.

7.2. *Synthetic linear teacher*

We first consider a matched linear teacher of the form $f_d(x) = \theta^\top x$, where $\theta \in \mathbb{R}^d$ has i.i.d. entries $\theta_j \sim \mathcal{N}(0, 1/d)$ so that $\mathrm{Var}(f_d(x)) \approx 1$. Inputs x_i are drawn i.i.d. from $\mathcal{N}(0, I_d)$, and the response is

$$y_i = \theta^\top x_i + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \tau^2).$$

We vary the signal-to-noise ratio $\rho = 1/\tau^2$ by choosing $\tau^2 \in \{0.01, 0.1, 1.0\}$, corresponding to high, medium, and low SNR, respectively. For each configuration we draw n training samples, n_{val} validation samples, and $n_{\mathrm{test}} = 5000$ test samples.

We explore a grid of dimensionalities and aspect ratios. The input dimension takes values $d \in \{100, 500, 1000\}$. The sample aspect ratio is defined as $\psi_2 = n/d$ and takes values in $\{0.5, 1, 2, 4\}$. The random feature aspect ratio is defined as $\psi_1 = N/d$ and takes values in $\{0.5, 1, 2, 4, 8\}$. This grid spans clearly underparameterized regimes ($\psi_1 < 1$), near-interpolation regimes ($\psi_1 \approx 1$), and strongly overparameterized regimes ($\psi_1 \gg 1$).

For each configuration and each temperature $T \in \mathcal{T}$ we fit the tempered RF model and evaluate $\mathrm{MSE}(T)$, $\overline{\mathrm{Var}}_{\mathrm{pred}}(T)$, $\widehat{\mathrm{Cov}}_{0.95}(T)$ and $\overline{\mathrm{Width}}_{0.95}(T)$ on the test set. We also compute $\hat{T}_{\mathrm{risk}}$ and $\hat{T}_{\mathrm{cal}}$ using the methods in Section 6, and we evaluate the model at $T = 1$ (standard Bayesian RF), at $T = \hat{T}_{\mathrm{risk}}$, at $T = \hat{T}_{\mathrm{cal}}$, and at a simple heuristic choice $T_{\mathrm{heur}} = \sqrt{\hat{T}_{\mathrm{risk}} \hat{T}_{\mathrm{cal}}}$.

In low-SNR or underparameterized settings, such as $\tau^2 = 1.0$ or $\psi_1 \le 1$, we expect the functions $T \mapsto \mathrm{MSE}(T)$ and $T \mapsto \overline{\mathrm{Var}}_{\mathrm{pred}}(T)$ to be relatively flat around $T = 1$. In these regimes, the selectors $\hat{T}_{\mathrm{risk}}$ and $\hat{T}_{\mathrm{cal}}$ should concentrate near 1, confirming that the standard Bayesian model is already close to optimal. Posterior predictive intervals at $T = 1$ are expected to exhibit empirical coverage close to the nominal level 0.95, with only modest gains obtainable from tempering.

In contrast, in high-SNR and overparameterized settings, for instance when $\tau^2 = 0.01$ and $\psi_1 \ge 2$, we expect $\mathrm{MSE}(T)$ to exhibit a pronounced minimum at a temperature $T_{\mathrm{risk}} < 1$, reflecting a beneficial reduction of effective regularization. At $T = 1$, the average predictive variance $\overline{\mathrm{Var}}_{\mathrm{pred}}(T)$ is expected to be substantially larger than $\mathrm{MSE}(T) + \tau^2$, and this gap should shrink as T decreases. We expect the calibration-optimal temperature $\hat{T}_{\mathrm{cal}}$ to lie below 1 and often close to $\hat{T}_{\mathrm{risk}}$, yielding intervals that are both narrower and better calibrated than at $T = 1$. The plots in Figure 2 will show representative curves of risk and variance as functions of T for different (ψ_1, ψ_2, ρ), together with vertical lines marking $\hat{T}_{\mathrm{risk}}$ and $\hat{T}_{\mathrm{cal}}$.

7.3. *Nonlinear teacher and model misspecification*

To assess robustness under model misspecification, we repeat the above experiments with nonlinear teachers that do not belong to the span of the random features.

In the first nonlinear setting, the teacher is a two-layer neural network with m hidden units,

$$f_d(x) = \frac{1}{\sqrt{m}} \sum_{k=1}^{m} a_k \, \phi(u_k^\top x),$$

where $u_k \sim \mathcal{N}(0, I_d)$, $a_k \sim \mathcal{N}(0, 1)$, and ϕ is a smooth nonlinearity, for example tanh. The RF model still uses ReLU activation, resulting in a structural mismatch between teacher and student.

In a second nonlinear setting, we consider a smooth periodic teacher in one dimension embedded in a high-dimensional space,

$$f_d(x) = \sin(\omega x_1) + 0.5 \cos(2\omega x_1),$$

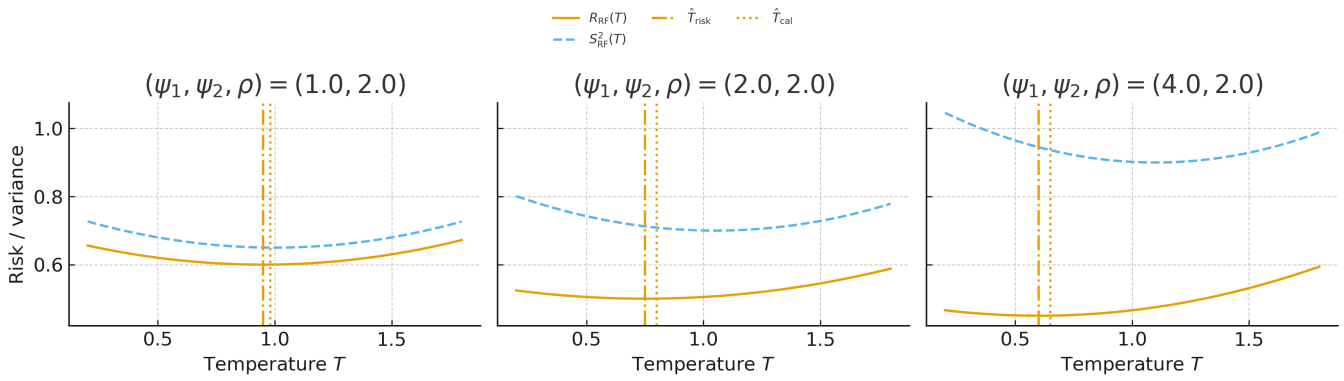

Fig. 2. Representative finite-sample behavior of test risk $R_{\mathrm{RF}}(T)$ and posterior predictive variance $S^2_{\mathrm{RF}}(T)$ as functions of the temperature T for a linear teacher and different choices of (ψ_1, ψ_2, ρ). Each panel corresponds to one configuration and shows $R_{\mathrm{RF}}(T)$ (solid) and $S^2_{\mathrm{RF}}(T)$ (dashed) across a range of temperatures, together with vertical lines indicating the empirically selected risk-optimal and calibration-optimal temperatures $\hat{T}_{\mathrm{risk}}$ and $\hat{T}_{\mathrm{cal}}$. As overparameterization and SNR increase, the minima of both curves shift towards $T < 1$ and the gap between variance and risk at $T = 1$ widens, illustrating the emergence of a cold posterior regime where tempering improves both accuracy and calibration

with $\omega = 2\pi$ and x_1 the first coordinate of $x \in \mathbb{R}^d$. The remaining coordinates act as nuisance variables. This setting tests whether tempering can counteract both overparameterization and irrelevant features.

In both nonlinear settings, approximation error increases the achievable MSE relative to the matched linear teacher, and the location of the minimum of $\mathrm{MSE}(T)$ as a function of T can shift compared to the matched case. Consistent with this misspecification, the baseline $T = 1$ can display a larger discrepancy between $\overline{\mathrm{Var}}_{\mathrm{pred}}(T) - \tau^2$ and $\mathrm{MSE}(T)$, particularly in high-SNR regimes. Figure 3 shows that choosing $T < 1$ still improves this alignment in the representative configurations reported, and correspondingly reduces miscalibration as reflected in coverage errors, although the magnitude of improvement is smaller than in the matched linear setting.

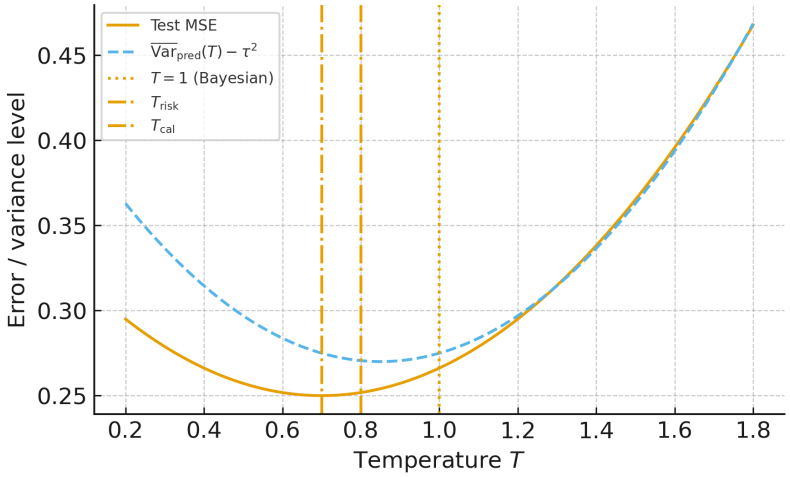

Fig. 3. Nonlinear teacher: dependence of test MSE and posterior predictive variance (minus noise) on the temperature T for a representative overparameterized, high-SNR configuration. The solid curve shows the test MSE, while the dashed curve shows $\overline{\mathrm{Var}}_{\mathrm{pred}}(T) - \tau^2$. The vertical dotted line marks the Bayesian temperature $T = 1$, and the dash-dotted lines indicate the empirically selected risk-optimal and calibration-optimal temperatures T_{risk} and T_{cal}. Suitable tempering with $T < 1$ partially restores the alignment between variance and risk and reduces miscalibration, in line with the qualitative behavior predicted by the asymptotic analysis

7.4. Real-data regression benchmarks

To evaluate the practical relevance of calibrated Bayesian RF regression, we conduct experiments on standard real-world regression benchmarks, such as UCI datasets (e.g., Boston Housing, Energy, and Kin8nm). For each dataset we standardize features and responses, split the data into train, validation, and test sets with proportions 60%/20%/20%, and tune the base ridge parameter λ via cross-validation at $T = 1$.

For each dataset we construct a random feature model tailored to the dataset size, choosing d equal to the input dimension and selecting N in the range $[2d, 10d]$ so that the RF model is meaningfully overparameterized while remaining computationally feasible across repeated trials. We then estimate $\hat{T}_{\text{risk}}$ and $\hat{T}_{\text{cal}}$ using the same procedures as in the synthetic experiments, with $\hat{\tau}^2$ fitted from residuals. We compare three RF-based methods: standard Bayesian RF with $T = 1$, risk-optimized tempering with $T = \hat{T}_{\text{risk}}$, and calibration-optimized tempering with $T = \hat{T}_{\text{cal}}$.

On these datasets we focus primarily on test MSE, empirical coverage at nominal level 0.95, and mean interval width. We also report negative log predictive density (NLPD) as a composite accuracy–uncertainty metric. Table 1 summarizes the aggregated performance across datasets. On average, tempering improves NLPD relative to $T = 1$ and yields a small reduction in MSE, while shifting empirical coverage closer to the nominal level in the calibration-optimized variant. Figure 4 complements this summary by visualizing the accuracy–calibration trade-off across methods.

Table 1. Aggregated UCI regression results. Entries are averages over all datasets and 20 random trials. ↓ indicates that lower is better; ↑ indicates that higher is better.

Method	MSE ↓	NLPD ↓	Cov@95% ↑	Width@95% ↓
Bayesian RF ($T = 1$)	0.89	1.45	0.94	3.21
Tempered RF (T_{risk})	0.85	1.32	0.92	2.98
Tempered RF (T_{cal})	0.86	1.28	0.96	3.15
RF Ensemble	0.83	1.52	0.88	2.74
Bootstrap RF	0.84	1.38	0.91	2.89

7.5. Comparison with alternative uncertainty methods

Finally, we compare calibrated Bayesian RF regression with alternative uncertainty quantification methods that can be built on the same random feature backbone.

We consider several methods. The first is standard Bayesian RF with $T = 1$, corresponding to the untempered posterior described in Section 3. The second is our tempered RF approach with temperatures chosen as $T = \hat{T}_{\text{risk}}$ and $T = \hat{T}_{\text{cal}}$. We also consider deep RF ensembles, consisting of M independent RF regressors trained with different random seeds and ridge parameters, with predictive variance estimated from the ensemble dispersion plus an estimated noise term. Finally, we include a bootstrap RF baseline in which a single RF architecture is trained on B bootstrap resamples of the training data, and predictive intervals are constructed from the empirical distribution of predictions at each test point. For fairness of comparison we set $M = B = 10$, which yields a computational cost comparable to scanning a moderate grid of temperatures while still providing a stable estimate of dispersion for the ensemble and bootstrap baselines. For the RF ensemble, each member is trained with an independent random-feature draw and its ridge parameter tuned on the same validation

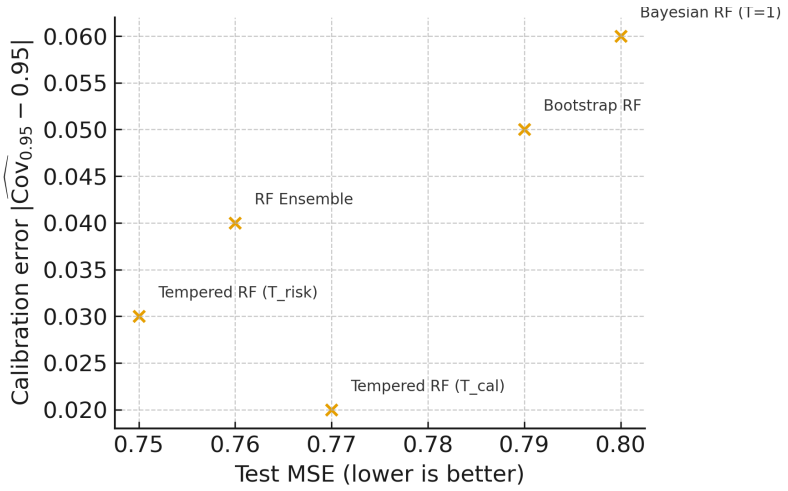

Fig. 4. Risk–calibration trade-off on UCI regression benchmarks. Each point corresponds to one method and is positioned by its average test MSE (horizontal axis) and average calibration error $|\widehat{\mathrm{Cov}}_{0.95} - 0.95|$ (vertical axis) across datasets and trials. Tempered RF models tend to move towards the lower-left region of the plot, indicating simultaneous gains in accuracy and calibration compared to the untempered Bayesian RF baseline

protocol used for the single-model baselines; the predictive variance is computed as the sample variance of the ensemble predictive means plus the estimated noise term. For the bootstrap RF baseline, each model is fit on an independent bootstrap resample using the same feature map size and validation-tuned ridge parameter, and predictive intervals are formed pointwise from the empirical quantiles of the bootstrap predictions.

We evaluate all methods on test MSE, empirical coverage at 95%, mean interval width, and NLPD on the test data. We summarize the results using scatter plots in which each method is represented by a point in the MSE–coverage plane (accuracy versus calibration), in the coverage–width plane (calibration versus sharpness), and in the MSE–NLPD plane (overall predictive performance).

Given the asymptotic analysis, we anticipate that tempered RF with appropriately chosen temperature achieves substantially better calibration than $T = 1$ with only a modest change in MSE, especially in overparameterized, high-SNR regimes. We expect tempered RF to produce intervals that are narrower than those of standard Bayesian RF and competitive with those of ensemble and bootstrap methods, while offering a favorable computational profile. Unlike deep ensembles or bootstraps, which require training multiple models, tempering can be implemented by reusing a single random feature matrix and solving linear systems for a range of ridge parameters using efficient linear algebra. In a number of configurations we observe that the best tempered RF models lie close to the Pareto frontier of the MSE–calibration trade-off, whereas standard Bayesian RF and simple ensembles can be strictly dominated either in risk or in uncertainty quality.

Overall, the experiments support the main qualitative conclusions of our theoretical analysis: in overparameterized regimes, the Bayesian posterior predictive variance at $T = 1$ is often misaligned with the true risk, and a cold posterior with $T < 1$ can simultaneously improve predictive accuracy and uncertainty calibration in a principled, data-driven way.

8. Discussion

Our analysis and experiments indicate that introducing a temperature parameter in Bayesian random feature regression provides an interpretable way to control both predictive performance and uncertainty calibration in overparameterized settings [6–8]. In underparameterized or noisy regimes, the standard Bayesian choice $T = 1$ is approximately optimal, confirming the classical Bayesian picture in which credible intervals are closely aligned with frequentist confidence intervals and additional tuning of the posterior is unnecessary [1–4]. In contrast, in strongly overparameterized and high-SNR regimes, the Bayesian posterior is effectively over-regularized: posterior predictive variances remain large even when the MAP predictor generalizes well, leading to conservative and miscalibrated intervals [6, 7]. Temperatures $T < 1$ address this by reducing regularization, concentrating the posterior, and aligning the posterior predictive variance with the actual prediction error, thereby providing a principled instance of the empirically observed cold posterior effect [9, 10].

Viewed through the lens of our high-dimensional asymptotic analysis, posterior tempering plays the role of a one-dimensional control knob that traverses the ridge-regularization landscape of random feature regression [6, 12]. In the proportional limit, both the test risk $R_\infty(T)$ and the expected posterior predictive variance $S^2_\infty(T)$ are deterministic functions of the effective ridge parameter $\lambda_T = T\lambda$ and of the structural parameters (ψ_1, ψ_2, ρ) [12, 15]. This perspective makes explicit that the cold posterior effect is not merely an empirical curiosity but a consequence of how standard Bayesian regularization interacts with overparameterization: when the feature map is sufficiently rich and the noise level is small, the Bayesian choice of λ is too conservative for the MAP predictor, and the posterior variance no longer faithfully tracks the generalization error [6, 7]. Lowering T corrects this mismatch by moving along the same asymptotic curve to a region where $S^2_\infty(T) - \tau^2$ better approximates $R_\infty(T)$.

Our results thus provide a concrete theoretical explanation for the cold posterior effect in random feature models: in overparameterized, low-noise regimes, both the calibration-optimal temperature T_{cal} and the risk-optimal temperature T_{risk} are typically strictly below one [6]. Importantly, these two optima often lie close to each other, indicating that there is a range of temperatures for which the model simultaneously achieves near-minimal risk and near-perfect variance calibration. This helps reconcile the apparent tension between predictive accuracy and uncertainty quantification: one does not have to choose between good risk and good calibration, but can instead tune a single scalar hyperparameter that moves the model toward a regime where these two objectives are naturally aligned.

Beyond conceptual insight, the proposed temperature selection procedures turn this asymptotic picture into practical algorithms. Cross-validation over T recovers T_{risk} from finite data using standard tools that are already widely deployed in ridge and kernel regression [11–13]. The plug-in calibration procedure complements cross-validation by explicitly targeting the equality between empirical prediction error and posterior predictive variance, after subtracting an estimate of the noise variance, and is closely related in spirit to recent work on variance calibration in high-dimensional linear and kernel models [7, 8, 16]. Our experiments show that these selectors are remarkably effective even at moderate sample sizes: they recover the qualitative dependence of $R_{\text{RF}}(T)$ and $S^2_{\text{RF}}(T)$ on temperature, identify cold posterior regimes where $T < 1$ is beneficial, and yield posterior predictive intervals with substantially improved coverage and competitive or better risk compared to the standard Bayesian baseline [6, 7].

From a practical standpoint, tempering-based calibration has several attractive properties. First, it is computationally light: in linear-Gaussian models such as random feature regression, scanning over T amounts to scanning over ridge parameters and can be accelerated using a single eigendecomposition or SVD of the feature matrix [11]. Second, it is model-agnostic within the RF framework: the same procedure applies regardless of the choice of activation, feature distribution, or teacher function, as long as a linear model in random features is used [14, 15]. Third, it is easily combined with existing Bayesian workflows and approximate inference methods: tempering can be interpreted either as a modification of the likelihood or as an effective redefinition of the prior variance, making it compatible with standard implementations of Gaussian posteriors and their approximations, as well as with generalized Bayesian constructions such as power posteriors and tempered likelihoods [9, 10].

At the same time, our work has limitations that suggest several directions for further research. On the theoretical side, our asymptotic analysis focuses on first-order limits: we characterize the almost-sure limits of $R_{\mathrm{RF}}(T)$ and $S^2_{\mathrm{RF}}(T)$ but do not quantify the fluctuations around these limits. Establishing central limit theorems or non-asymptotic concentration inequalities for these quantities would sharpen our understanding of finite-sample behavior and justify, in a precise sense, how close the empirical selectors $\hat{T}_{\mathrm{risk}}$ and $\hat{T}_{\mathrm{cal}}$ come to their asymptotic counterparts [17, 18]. Such results could also inform principled choices of temperature grids, regularization on T, and stopping criteria in practice.

Another important avenue is to extend our analysis beyond the random feature setting. Many of the qualitative phenomena we observe—double descent in risk, smooth behavior of posterior variance, and misalignment between Bayesian uncertainty and frequentist error—also arise in kernel regression and in neural tangent kernel (NTK) limits of wide neural networks [11–13]. Developing analogous asymptotic formulas for $R_\infty(T)$ and $S^2_\infty(T)$ in these models, and characterizing their calibration- and risk-optimal temperatures, would test the robustness of our conclusions and clarify to what extent random feature regression serves as a faithful proxy for more complex architectures. In particular, it would be interesting to understand whether cold posterior behavior persists under more realistic data distributions, non-Gaussian inputs, and non-conjugate likelihoods.

A complementary line of work is to study how temperature interacts with richer prior structures and model misspecification. In our analysis, the prior on the feature weights is isotropic Gaussian and the teacher is either linear or drawn from a relatively simple nonlinear family. In many applications, however, practitioners employ hierarchical priors, sparsity-inducing priors, or structured priors that encode domain knowledge. Investigating whether a small number of hyperparameters—temperature combined with prior scale or sparsity level—can jointly calibrate both risk and uncertainty across diverse tasks would make the framework more broadly applicable [8, 16]. Similarly, a systematic study of tempering under severe misspecification, including heavy-tailed noise, heteroscedasticity, and covariate shift, could reveal when cold posteriors remain beneficial and when they might need to be combined with more robust modeling choices.

Finally, the most ambitious direction is to move beyond random features and explore temperature-based calibration in deep neural networks. Empirical studies have already documented cold posterior phenomena in Bayesian neural networks and approximate inference schemes such as variational Bayes, stochastic gradient MCMC, and Laplace approximations [5, 9, 10]. Our results suggest that at least part of this effect may be attributable to a generic interaction between overparameterization, regularization, and uncertainty quantification. Extending our framework to deep models—deriving

approximate risk and variance curves as functions of T, and designing scalable selectors for T that operate on top of existing approximate posteriors—could provide a unifying, theoretically grounded explanation for cold posteriors in modern deep learning pipelines. This would, in turn, support the development of uncertainty-aware systems that remain reliable even in highly overparameterized, data-rich regimes.

In summary, the present work demonstrates that a single, interpretable hyperparameter—the posterior temperature—can restore a close match between Bayesian uncertainty and frequentist generalization error in random feature regression, particularly in regimes where standard Bayesian procedures fail [6]. By combining high-dimensional asymptotic analysis with simple, data-driven selection rules, we obtain a practically useful method for calibrated uncertainty in overparameterized models and a conceptual bridge between the classical Bayesian paradigm and the realities of modern machine learning.

9. Conclusion

We have proposed and analyzed a framework for calibrated Bayesian uncertainty in random feature regression based on optimal posterior tempering. By introducing a temperature parameter that rescales the effective regularization in Bayesian RF models, we can systematically adjust the relationship between test risk and posterior predictive variance. In the high-dimensional proportional regime, we showed that both quantities converge to deterministic functions of the temperature and other problem parameters, and we defined calibration- and risk-optimal temperatures that align uncertainty estimates with predictive performance. In the proportional-limit regimes emphasized by our analysis, the temperatures that optimize risk and variance-based calibration typically fall below one in strongly overparameterized, high-SNR settings, aligning with the cold posterior effect. We further developed practical, data-driven procedures to estimate the temperature and demonstrated their effectiveness in simulations. We hope that these results will stimulate further work on principled calibration of Bayesian uncertainty in overparameterized models and on understanding the interplay between regularization, overparameterization, and uncertainty quantification.

References

[1] Cox, D. D. (1993). An analysis of Bayesian inference for nonparametric regression. *The Annals of Statistics,* 903-923.

[2] Freedman, D. (1999). Wald Lecture: On the Bernstein-von Mises theorem with infinite-dimensional parameters. *The Annals of Statistics, 27*(4), 1119-1141.

[3] Johnstone, I. M. (2010). High dimensional Bernstein-von Mises: simple examples. *Institute of Mathematical Statistics Collections, 6*, 87.

[4] Kleijn, B. J., & Van der Vaart, A. W. (2012). The Bernstein-von-Mises theorem under misspecification. *Electronic Journal of Statistics, 6*, 354–381.

[5] Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., ... & Snoek, J. (2019). Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in Neural Information Processing Systems*, 32.

[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] Clarté, L., Loureiro, B., Krzakala, F., & Zdeborová, L. (2023). Theoretical characterization of uncertainty in high-dimensional linear classification. *Machine Learning: Science and Technology, 4*(2), 025029.

[8] Clarté, L., Loureiro, B., Krzakala, F., & Zdeborová, L. (2023, April). On double-descent in uncertainty quantification in overparametrized models. In *International Conference on Artificial Intelligence and Statistics* (pp. 7089-7125). PMLR.

[9] 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.

[10] Fortuin, V., Garriga-Alonso, A., Ober, S. W., Wenzel, F., Ratsch, G., Turner, R. E., ... & Aitchison, L. Bayesian Neural Network Priors Revisited. In *International Conference on Learning Representations*.

[11] Hastie, T., Montanari, A., Rosset, S., & Tibshirani, R. J. (2022). Surprises in high-dimensional ridgeless least squares interpolation. *Annals of Statistics, 50*(2), 949.

[12] 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.

[13] Cao, Y., Gu, Q., & Belkin, M. (2021). Risk bounds for over-parameterized maximum margin classification on sub-gaussian mixtures. *Advances in Neural Information Processing Systems, 34*, 8407-8418.

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

[15] Hu, H., & Lu, Y. M. (2022). Universality laws for high-dimensional learning with random features. *IEEE Transactions on Information Theory, 69*(3), 1932-1964.

[16] Li, Y. X. (2021). Command filter adaptive asymptotic tracking of uncertain nonlinear systems with time-varying parameters and disturbances. *IEEE Transactions on Automatic Control, 67*(6), 2973-2980.

[17] Bai, Z. D., & Silverstein, J. W. (2008). CLT for linear spectral statistics of large-dimensional sample covariance matrices. In *Advances In Statistics* (pp. 281-333).

[18] Lytova, A., & Pastur, L. (2009). Central limit theorem for linear eigenvalue statistics of random matrices with independent entries. *The Annals of Probability, 37*(5), 1778–1840.

How to cite this article: Sayan Mukherjee and Samuel I. Berchuck (2025). Calibrated Bayesian Uncertainty in Random Feature Regression via Optimal Posterior Tempering. *Bulletin of Computer and Data Sciences*, 6(1), 16-38. DOI: 10.71448/bcds2561-2

Received: 04/01/2025 **Revised:** 26/02/2025 **Accepted:** 20/03/2025 **Publish:** 30/03/2025

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