MONOTONICITY AND DOUBLE DESCENT IN UNCERTAINTY ESTIMATION WITH GAUSSIAN PROCESSES

Anonymous authors
Paper under double-blind review

ABSTRACT

The quality of many modern machine learning models improves as model complexity increases, an effect that has been quantified—for predictive performance—with the non-monotonic double descent learning curve. Here, we address the overarching question: is there an analogous theory of double descent for models which estimate uncertainty? We provide a partially affirmative and partially negative answer in the setting of Gaussian processes (GP). Under standard assumptions, we prove that higher model quality for optimally-tuned GPs (including uncertainty prediction) under marginal likelihood is realized for larger input dimensions, and therefore exhibits a monotone error curve. After showing that marginal likelihood does not naturally exhibit double descent in the input dimension, we highlight related forms of posterior predictive loss that do exhibit non-monotonicity. Finally, we verify empirically that our results hold for real data, beyond our considered assumptions, and we explore consequences involving synthetic covariates.

1 INTRODUCTION

With the recent success of overparameterized and nonparametric models for many predictive tasks in machine learning (ML), the development of the corresponding uncertainty quantification (UQ) has unsurprisingly become a topic of significant interest. Naïve approaches for forward propagation of error and other methods for inverse uncertainty problems typically apply Monte Carlo methods under a Bayesian framework (Zhang, 2021). However, the large-scale nature of many ML problems of interest results in significant computational challenges. One of the most successful approaches for solving inverse uncertainty problems is the use of *Gaussian processes* (GP) (Williams & Rasmussen, 2006). This is now frequently used for many predictive tasks, including time-series analysis (Roberts et al., 2013) and classification (Williams & Barber, 1998; Williams & Rasmussen, 2006). GPs are also fundamental for Bayesian optimization (Hebbal et al., 2019; Frazier, 2018), although extending Bayesian optimization into high-dimensional settings remains challenging (Binois & Wycoff, 2021).

Although the theoretical understanding of the predictive capacity of high-dimensional ML models continues to advance rapidly, a parallel rigorous theory for UQ is comparatively lagging. The prominent heuristic in modern ML that larger models will typically perform better has become almost axiomatic. However, it is only more recently that this heuristic has become represented in the theory through the characterisation of benign overfitting (Bartlett et al., 2020). In particular, the *double descent* curve extends the bias-variance tradeoff curve to account for improving performance with higher model complexity (Belkin et al., 2019; Wang et al., 2021; Derezinski et al., 2020b) (see Figure 1(right)). Typically, these arguments involve applications of random matrix theory (Edelman & Rao, 2005; Paul & Aue, 2014), notably the Marchenko-Pastur law, concerning limits of spectral distributions under large data/large dimension regimes.

While the predictive performance of ML models can improve as model size increases, it is not clear whether or not the same is true for predictions of model uncertainty. Several common measures of model quality which incorporate inverse uncertainty quantification are Bayesian in nature, the most prominent of which are the *marginal likelihood* and various forms of posterior predictive loss. It is well-known that Bayesian methods can perform well in high dimensions (De Roos et al., 2021), even outperforming their low-dimensional counterparts when properly tuned (Wilson & Izmailov, 2020). To close this theory-practice gap, an analogous formulation of double descent curves in the setting of uncertainty quantification is desired. Marginal likelihood and posterior distributions are often

Performance Metric	Error Curve	Optimal γ
Marginal Likelihood / Free Energy (3)	Monotone (Thm. 1)	eqn. (5)
Posterior Predictive L^2 Loss (1)	Double Descent (Prop. 1)	0
Posterior Predictive NLL (2)	Double Descent (Prop. 1)	$\mathbb{E}\|\bar{f}(x) - y\|^2$

Table 1: Behavior of UQ performance metrics and optimal posterior temperature γ.

intractable for arbitrary models (e.g., Bayesian neural networks (Goan & Fookes, 2020)). However, their explicit forms are well known for GPs (Williams & Rasmussen, 2006).

GPs are nonparametric, and most of the kernels used in practice induce infinite-dimensional feature spaces, so model complexity can be difficult to quantify (although some notions of kernel dimension have been proposed (Zhang, 2005; Alaoui & Mahoney, 2015)). Nevertheless, it is generally expected that accurately fitting a GP to data lying in higher-dimensional spaces requires training on a larger dataset. This *curse of dimensionality* has been justified using error estimates (von Luxburg & Bousquet, 2004), and verified empirically (Spigler et al., 2020). However, under appropriate setups, predictive performance has been demonstrated to *improve* with larger input dimension (Liu et al., 2021). Here, we consider whether the same is true for the marginal likelihood and posterior predictive metrics. Our main results (see Theorem 1 and Proposition 1) are summarized as follows.

- ***Monotonicity****: For two optimally regularized scalar GPs, both fit to a sufficiently large set of iid normalized and whitened input-output pairs, the better performing model under marginal likelihood is the one with larger input dimension.*
- ***Double Descent****: For sufficiently small temperatures, GP posterior predictive metrics exhibit double descent if and only if the mean squared error for the corresponding kernel regression task exhibits double descent (see Liang & Rakhlin (2020); Liu et al. (2021) for sufficient conditions).*

Figure 1 illustrates characteristics of monotone and double descent error curves. Along the way, we identify optimal choices of temperature (which can be interpreted as noise in the data) under a tempered posterior setup — see Table 1 for a summary. Our results highlight that the common curse of dimensionality heuristic can be bypassed through an empirical Bayes procedure. Furthermore, the performance of optimally regularized GPs (under several metrics), can be improved with additional covariates (including synthetic ones). Our theory is supported by experiments performed on real large datasets. Additional experiments, including the effect of ill-conditioned inputs, alternative data distributions, and choice of underlying kernel, are conducted in Appendix A. Details of the setup for each experiment are listed in Appendix G.

2 Background

2.1 Gaussian Processes

A *Gaussian process* is a stochastic process f on $\mathbb{R}^d$ such that for any set of points $x_1, \ldots, x_k \in \mathbb{R}^d$, $(f(x_1), \ldots, f(x_k))$ is distributed as a multivariate Gaussian random vector (Williams & Rasmussen, 2006, §2.2). Gaussian processes are completely determined by their *mean* and *covariance functions*: if for any $x, x' \in \mathbb{R}^d$, $\mathbb{E}f(x) = m(x)$ and $\mathrm{Cov}(f(x), f(x')) = k(x, x')$, then we say that $f \sim \mathcal{GP}(m, k)$. Inference for GPs is informed by Bayes' rule: letting $(X_i, Y_i)_{i=1}^n$ denote a collection of iid input-output pairs, we impose the assumption that $Y_i = f(X_i) + \epsilon_i$ where each $\epsilon_i \sim \mathcal{N}(0, \gamma)$,

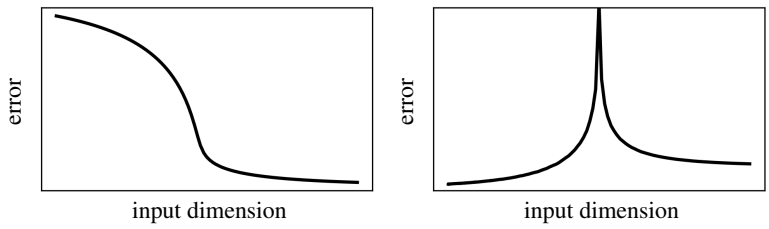

Figure 1: Illustrations of monotone (left) and double descent (right) error curves.

yielding a Gaussian likelihood $p(Y|f,X) = (2\pi\gamma)^{-n/2}\exp(-\frac{1}{2\gamma}\|Y - f(X)\|^2)$. The parameter γ is the *temperature* of the model, and dictates the perceived accuracy of the labels. For example, taking $\gamma \to 0^+$ considers a model where the labels are noise-free.

For the prior, we assume that $f \sim \mathcal{GP}(0, \lambda^{-1}k)$ for a fixed covariance kernel k and regularization parameter $\lambda > 0$. While other mean functions m can be considered, in the sequel we will consider the case where $m \equiv 0$. Indeed, if $m \neq 0$, then one can instead consider $\tilde{Y}_i = Y_i - m(X_i)$, so that $\tilde{Y}_i = \tilde{f}(X_i) + \epsilon_i$ and the corresponding prior for $\tilde{f}$ is zero-mean. The Gram matrix $K_X \in \mathbb{R}^{n\times n}$ for X has elements $K_X^{ij} = k(X_i, X_j)$. Let $\boldsymbol{x} = (x_1, \ldots, x_m)$ denote a collection of N points in $\mathbb{R}^d$, $f(\boldsymbol{x}) = (f(x_1), \ldots, f(x_m))$ and denote by $K_{\boldsymbol{x}} \in \mathbb{R}^{m\times m}$ and $k_{\boldsymbol{x}} \in \mathbb{R}^{n\times m}$ the matrices with elements $K_{\boldsymbol{x}}^{ij} = k(x_i, x_j)$ and $k_{\boldsymbol{x}}^{ij} = k(X_i, x_j)$.

Given this setup, we are interested in several metrics which quantify the uncertainty of the model. The **posterior predictive distribution** (PPD) of $f(\boldsymbol{x})$ given X, Y is (Williams & Rasmussen, 2006, pg. 16)

$$f(\boldsymbol{x})|X, Y \sim \mathcal{N}(\bar{f}(\boldsymbol{x}), \lambda^{-1}\Sigma(\boldsymbol{x})),$$

where $\bar{f}(\boldsymbol{x}) = k_{\boldsymbol{x}}^\top(K_X + \lambda\gamma I)^{-1}Y$ and $\Sigma(\boldsymbol{x}) = K_{\boldsymbol{x}} - k_{\boldsymbol{x}}^\top(K_X + \lambda\gamma I)^{-1}k_{\boldsymbol{x}}$. This defines a posterior predictive distribution ρ^γ on the GP f given X, Y (so $f \mid X, Y \sim \rho^\gamma$). If we let $\boldsymbol{y} = (y_1, \ldots, y_m)$ denote a collection of m associated *test labels* corresponding to our test data $\boldsymbol{x}$, the **posterior predictive** L^2 **loss** (PPL2) is the quantity

$$\ell(\boldsymbol{x}, \boldsymbol{y}) \coloneqq \mathbb{E}_{f\sim\rho^\gamma}\|f(\boldsymbol{x}) - \boldsymbol{y}\|^2 = \|\bar{f}(\boldsymbol{x}) - \boldsymbol{y}\|^2 + \lambda^{-1}\mathrm{tr}(\Sigma(\boldsymbol{x})). \tag{1}$$

Closely related is the **posterior predictive negative log-likelihood** (PPNLL), given by

$$L(\boldsymbol{x}, \boldsymbol{y}|X, Y) \coloneqq -\mathbb{E}_{f\sim\rho^\gamma}\log p(\boldsymbol{y}|f, \boldsymbol{x}) = \tfrac{1}{2\gamma}\|\bar{f}(\boldsymbol{x}) - \boldsymbol{y}\|^2 + \tfrac{1}{2\lambda\gamma}\mathrm{tr}(\Sigma(\boldsymbol{x})) + \tfrac{m}{2}\log(2\pi\gamma). \tag{2}$$

2.2 Marginal Likelihood

The fundamental measure of model performance in Bayesian statistics is the *marginal likelihood* (also known as the *partition function* in statistical mechanics). Integrating the likelihood over the prior distribution π provides a probability density of data under the prescribed model. Evaluating this density at the training data gives an indication of model suitability before posterior inference. Hence, the marginal likelihood is $\mathcal{Z}_n = \mathbb{E}_{f\sim\pi}p(Y|f, X)$. A larger marginal likelihood is typically understood as an indication of superior model quality. The **Bayes free energy** $\mathcal{F}_n = -\log\mathcal{Z}_n$ is interpreted as an analogue of the test error, where smaller $\mathcal{F}_n$ is desired.

The **marginal likelihood for a Gaussian process** is straightforward to compute: since $Y_i = f(X_i) + \epsilon_i$ under the likelihood, and $(f(X_i))_{i=1}^n \sim \mathcal{N}(0, \lambda^{-1}K_X)$ under the GP prior $\pi = \mathcal{GP}(0, \lambda^{-1}k)$, we have $Y_i|X \sim \mathcal{N}(0, \lambda^{-1}K_X + \gamma I)$, and hence the Bayes free energy is given by (Williams & Rasmussen, 2006, eqn. (2.30))

$$\mathcal{F}_n^\gamma = \tfrac{1}{2}\lambda Y^\top(K_X + \lambda\gamma I)^{-1}Y + \tfrac{1}{2}\log\det(K_X + \lambda\gamma I) - \tfrac{n}{2}\log\left(\tfrac{\lambda}{2\pi}\right). \tag{3}$$

In practice, the hyperparameters λ, γ are often tuned to minimize the Bayes free energy. This is an *empirical Bayes procedure*, and typically achieves excellent results (Krivoruchko & Gribov, 2019).

The PPNLL can be linked to the marginal likelihood through cross-validation measures. Let I be uniform on $\{1, \ldots, k\}$ and let $\mathcal{T}$ be a random choice of k indices from $\{1, \ldots, n\}$ (the *test set*). Let $\bar{\mathcal{T}} = \{1, \ldots, n\}\backslash\mathcal{T}$ denote the corresponding *training set*. The leave-k-out cross-validation score under the PPNLL is defined by $S_k(X, Y) = \mathbb{E}L(X_{\mathcal{T}_I}, Y_{\mathcal{T}_I}|X_{\bar{\mathcal{T}}}, Y_{\bar{\mathcal{T}}})$. The Bayes free energy is the sum of all leave-k-out cross-validation scores (Fong & Holmes, 2020), that is $\mathcal{F}_n^\gamma = \sum_{k=1}^n S_k(X, Y)$. Therefore, the **mean Bayes free energy** (or mean free energy for brevity) $n^{-1}\mathcal{F}_n^\gamma$ can be interpreted as the average cross-validation score. Similar connections can also be formulated in the PAC-Bayes framework (Germain et al., 2016).

2.3 Bayesian Linear Regression

One of the most important special cases of GP regression is *Bayesian linear regression*, obtained by taking $k_{\text{lin}}(x, x') = x^\top x'$. As a special case of GPs, our results apply to Bayesian linear regression, directly extending double descent analysis into the Bayesian setting. By Mercer's Theorem

(Williams & Rasmussen, 2006, §4.3.1), a realization of a GP f has a series expansion in terms of the eigenfunctions of the kernel k. As the eigenfunctions of k_{lin} are linear, $f \sim \mathcal{GP}(0, \lambda^{-1} k_{\text{lin}})$ if and only if

$$f(x) = w^\top x, \qquad w \sim \mathcal{N}(0, \lambda^{-1}).$$

More generally, if $\phi : \mathbb{R}^d \to \mathbb{R}^p$ is a finite-dimensional feature map, then $f(x) = w^\top \phi(x)$, $w \sim \mathcal{N}(0, \lambda^{-1})$ is a GP with covariance kernel $k_\phi(x, y) = \phi(x)^\top \phi(y)$. This is the weight-space interpretation of Gaussian processes. In this setting, the posterior distribution over the weights satisfies $\rho^\gamma(w) = p(w|X, Y) \propto \exp(-\frac{1}{2\gamma}\|Y - \phi(X)w\|^2 - \frac{\lambda}{2}\|w\|^2)$ and the marginal likelihood becomes

$$\mathcal{Z}_n^\gamma = \int_{\mathbb{R}^p} p(Y|X, w)\pi(w)\mathrm{d}w = \frac{\lambda^{d/2}}{\gamma^{n/2}(2\pi)^{\frac{1}{2}(n+d)}} \int_{\mathbb{R}^p} e^{-\frac{1}{2\gamma}\|Y-\phi(X)w\|^2} e^{-\frac{\lambda}{2}\|w\|^2} \mathrm{d}w, \tag{4}$$

where $\phi(X) = (\phi(X_i))_{i=1}^n \in \mathbb{R}^{n \times p}$. Under this interpretation, the role of λ as a regularization parameter is clear. Note also that if $\lambda = \mu/\gamma$ for some $\mu > 0$, then the posterior $\rho^\gamma(w)$ depends on γ as $(\rho^1(w))^{1/\gamma}$ (excluding normalizing constants). This is called a *tempered posterior*; if $\gamma < 1$, the posterior is *cold*, and it is *warm* whenever $\gamma > 1$.

3 Related Work

Double Descent and Multiple Descent. *Double descent* is an observed phenomenon in the error curves of kernel regression, where the classical (U-shaped) bias-variance tradeoff in underparameterized regimes is accompanied by a curious monotone improvement in test error as the ratio c of the number of datapoints to the ambient data dimension increases beyond $c = 1$. The term was popularized in Belkin et al. (2018b; 2019). However, it had been observed in earlier reports (Dobriban & Wager, 2018; Loog et al., 2020), and the existence of such non-monotonic behavior as a function of system control parameters should not be unexpected, given general considerations about different phases of learning that are well-known from the statistical mechanics of learning (Engel & den Broeck, 2001; Martin & Mahoney, 2017). An early precursor to double descent analysis came in the form of the *Stein effect*, which establishes uniformly reduced risk when some degree of regularisation is added (Strawderman, 2021). Stein effects have been established for kernel regression in Muandet et al. (2014); Chang et al. (2017). Subsequent theoretical developments proved the existence of double descent error curves on various forms of linear regression (Bartlett et al., 2020; Tsigler & Bartlett, 2020; Hastie et al., 2022; Muthukumar et al., 2020), random features models (Liao et al., 2020; Holzmüller, 2020), kernel regression (Liang & Rakhlin, 2020; Liu et al., 2021), two-layer neural networks (Mei & Montanari, 2022), and classification tasks (Frei et al., 2022; Wang et al., 2021). For non-asymptotic results, subgaussian data is commonly assumed, yet other data distributions have also been considered (Derezinski et al., 2020b). Double descent error curves have also been observed in nearest neighbor models (Belkin et al., 2018a), decision trees (Belkin et al., 2019), and state-of-the-art neural networks (Nakkiran et al., 2021). More recent developments have identified a large number of possible curves in kernel regression (Liu et al., 2021), including triple descent (Adlam & Pennington, 2020; d'Ascoli et al., 2020) and multiple descent for related volume-based metrics (Derezinski et al., 2020a). Similar to our results, an optimal choice of regularization parameter can negate the double descent singularity and result in a monotone error curve (Liu et al., 2021; Nakkiran et al., 2020; Wu & Xu, 2020). While there does not appear to be clear consensus on a *precise* definition of "double descent," for our purposes, we say that an error curve $E(t)$ exhibits double descent if it contains a single global maximum away from zero at t^*, and decreases monotonically thereafter. This encompasses double descent as it appears in the works above, while excluding some misspecification settings and forms of multiple descent.

Learning Curves for Gaussian Processes. The study of error curves for GPs under posterior predictive losses has a long history (see Williams & Rasmussen (2006, §7.3) and Viering & Loog (2021)). However, most results focus on rates of convergence of posterior predictive loss in the large data regime $n \to \infty$. The resulting error curve is called a *learning curve*, as it tracks how fast the model learns with more data (Sollich, 1998; Sollich & Halees, 2002; Le Gratiet & Garnier, 2015). Of particular note are classical upper and lower bounds on posterior predictive loss (Opper & Vivarelli, 1998; Sollich & Halees, 2002; Williams & Vivarelli, 2000), which are similar in form to counterparts in the double descent literature (Holzmüller, 2020). For example, some upper bounds have been obtained with respect to forms of *effective dimension*, defined in terms of the Gram matrix

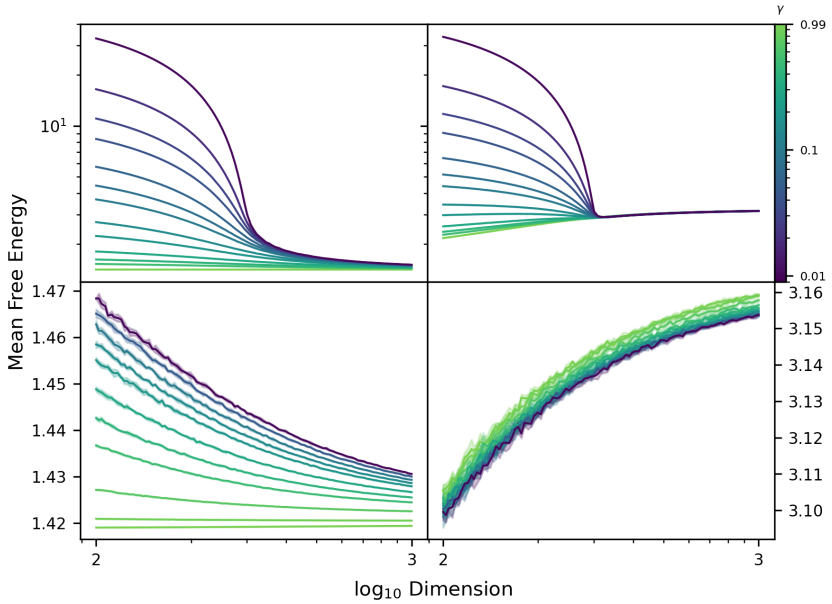

Figure 2: Error curves for mean Bayes free energy $n^{-1}\mathcal{F}_n^\gamma$ for **synthetic data** under linear (top) and Gaussian (bottom) kernels, with $\lambda = \lambda^*$ (left; **monotone decreasing**) and $\lambda = 0.01$ (right; **increases at higher input dimensions**).

(Zhang, 2005; Alaoui & Mahoney, 2015). Contraction rates in the posterior have also been examined (Lederer et al., 2019). In our work, we consider error curves over dimension rather than data, but we note that our techniques could also be used to study learning curves.

Cold Posteriors. Among the most surprising phenomena encountered in Bayesian deep learning is the *cold posterior effect* (CPE): the performance of Bayesian neural networks (BNNs) appears to improve for tempered posteriors when $\gamma \to 0^+$. This presents a challenge for uncertainty prediction: taking $\gamma \to 0^+$ concentrates the posterior around the *maximum a posteriori* (MAP) point estimator, and so the CPE implies that optimal performance is achieved when there is little or no predicted uncertainty. First observed in Wenzel et al. (2020), several authors have since sought to explain the phenomenon through data curation (Aitchison, 2020), data augmentation (Izmailov et al., 2021; Fortuin et al., 2021), and misspecified priors (Wenzel et al., 2020), although the CPE can still arise in isolation of each of these factors (Noci et al., 2021). While our setup is too simple to examine the CPE at large, we find some common forms of posterior predictive loss are optimized as $\gamma \to 0^+$.

4 Monotonicity in Bayes Free Energy

In this section, we investigate the behavior of the Bayes free energy using the explicit expression in (3). First, to facilitate our analysis, we require the following assumption on the kernel k.

Assumption. *The kernel k is formed by a function $\kappa : \mathbb{R} \to \mathbb{R}$ that is continuously differentiable on $(0, \infty)$ and is one of the following two types:*

(I) ***Inner product kernel:*** *$k(x, x') = \kappa(x^\top x'/d)$ for $x, x' \in \mathbb{R}^d$, where κ is three-times continuously differentiable in a neighbourhood of zero, with $\kappa'(0) > 0$. Let*

$$\alpha = \kappa'(0), \qquad \beta = \kappa(1) - \kappa(0) - \kappa'(0).$$

(II) ***Radial basis kernel:*** *$k(x, x') = \kappa(\|x - x'\|^2/d)$ for $x, x' \in \mathbb{R}^d$, where κ is three-times continuously differentiable on $(0, \infty)$, with $\kappa'(2) < 0$. Let*

$$\alpha = -2\kappa'(2), \qquad \beta = \kappa(0) + 2\kappa'(2) - \kappa(2).$$

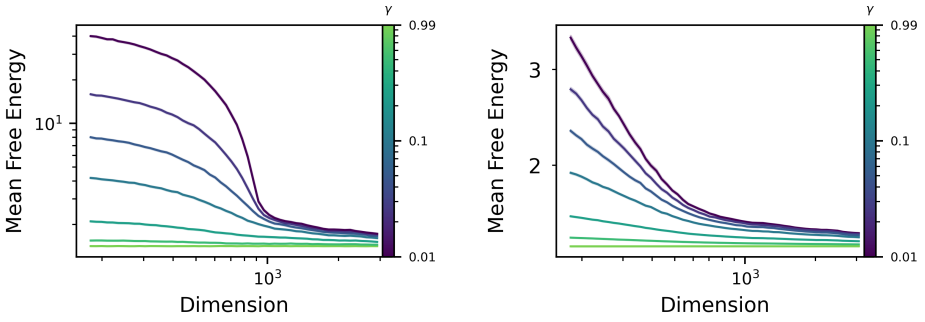

Figure 3: Error curves for mean Bayes free energy under the CIFAR10 dataset; linear (left) and Gaussian (right) kernels; $\lambda = \lambda^*$; **curves for real data match Figure 2 (left)**.

This assumption allows for many common covariance kernels used for GPs, including polynomial kernels $k(x, x') = (c + x^\top x'/d)^p$, the exponential kernel $k(x, x') = \exp(x^\top x'/d)$, the Gaussian kernel $k(x, x') = \exp(-\|x - x'\|^2/d)$, the multiquadric kernel $k(x, x') = (c + \|x - x'\|^2/d)^p$, the inverse multiquadric $k(x, x') = (c + \|x - x'\|^2/d)^{-p}$ kernels, and the Matérn kernels

$$k(x, x') \;=\; (2^{\nu-1}\Gamma(\nu))^{-1}\|x \;-\; x'\|^\nu K_\nu(\|x - x'\|)$$

(where K_ν is the Bessel-K function). Different bandwidths can also be incorporated through the choice of κ. However, it does exclude some of the more recent and sophisticated kernel families, e.g., neural tangent kernels. Due to a result of El Karoui (2010), the Gram matrices of kernels satisfying this assumption exhibit limiting spectral behavior reminiscent of that for the linear kernel, $k(x, x') = c + x^\top x'/d$. Roughly speaking, from the perspective of the marginal likelihood, we can treat GPs as Bayesian linear regression.

In line with previous work on double descent curves (Belkin et al., 2019), our objective is to investigate the behavior of the marginal likelihood with respect to model complexity, which is often given by the number of parameters in parametric settings (d'Ascoli et al., 2020; Derezinski et al., 2020b; Hastie et al., 2022)). GPs are non-parametric, and while notions of *effective dimension* do exist (Zhang, 2005; Alaoui & Mahoney, 2015), it is common to instead consider the input dimension in place of the number of parameters in this context (Liang & Rakhlin, 2020; Liu et al., 2021).

For our theory, we first consider the "best-case scenario," where the prior is perfectly specified and its mean function m is used to generate Y: $Y_i = m(X_i) + \epsilon_i$, where each ϵ_i is iid with zero mean and unit variance. By a change of variables, we can assume (without loss of generality) that $m \equiv 0$, so that $Y_i = \epsilon_i$, and is therefore independent of X. To apply the Marchenko-Pastur law from random matrix theory, we consider the large dataset – large input dimension limit, where n and d scale linearly so that $d/n \to c \in (0, \infty)$. The inputs are assumed to have been *whitened* and to be independent zero-mean random vectors with unit covariance. Under this limit, the sequence of mean Bayes entropies $n^{-1}\mathcal{F}_n^\gamma$, for each $n = 1, 2, \ldots$, converges in expectation over the training set to a quantity $\mathcal{F}_\infty^\gamma$ which is more convenient to study. Our main result is presented in Theorem 1; the proof is delayed to Supplementary Material.

Theorem 1 (Limiting Bayes Free Energy)**.** *Let $X_1, X_2, \ldots$ be independent and identically distributed zero-mean random vectors in $\mathbb{R}^d$ with unit covariance, satisfying $\mathbb{E}\|X_i\|^{5+\delta} < +\infty$ for some $\delta > 0$. For each $n = 1, 2, \ldots$, let $\mathcal{F}_n^\gamma$ denote (3) applied to $X = (X_i)_{i=1}^n$ and $Y = (Y_i)_{i=1}^n$, with each $Y_i \sim \mathcal{N}(0, 1)$. If $n, d \to \infty$ such that $d/n \to c \in (0, \infty)$, then*

$$\mathcal{F}_\infty^\gamma \coloneqq \lim_{n\to\infty} n^{-1}\mathbb{E}\mathcal{F}_n^\gamma,$$

is well-defined. In this case,

(a) If $\lambda = \mu/\gamma$ for some $\mu > 0$, there exists an optimal temperature γ^* which minimizes $\mathcal{F}_\infty^\gamma$, which is given by

$$\gamma^* = c - 1 - \tfrac{c}{\alpha}(\beta + \mu) + \sqrt{(1 + \tfrac{c}{\alpha}(\beta + \mu + \alpha))^2 - 4c}. \tag{5}$$

If the kernel k depends on λ such that α is constant in λ and $\beta = \beta_0\lambda$ for $\beta_0 \in [0, 1)$, then

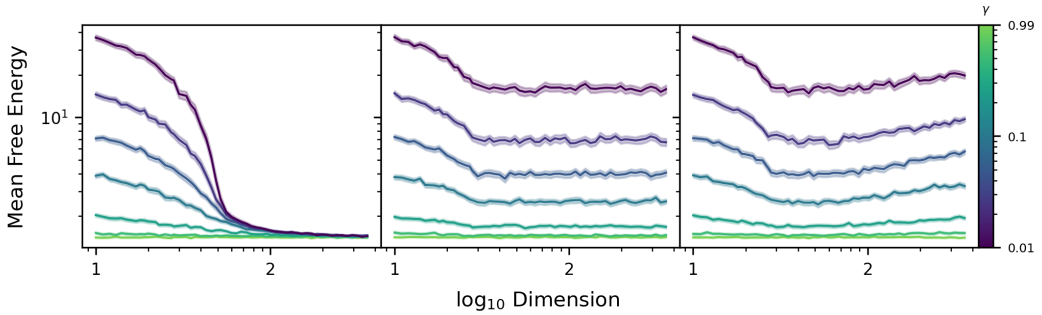

Figure 4: Error curves for mean Bayes free energy under real data with Gaussian (left); repeated data (center); and zeroed data (right), under the linear kernel and $\lambda = \lambda^*$. **Only adding synthetic non-zero iid covariates improves model performance.**

(b) If $\gamma \in (0, 1 - \beta_0)$, there exists a unique optimal $\lambda^* > 0$ minimizing $\mathcal{F}_\infty^\gamma$ satisfying

$$\lambda^* = \frac{\alpha[(c+1)(\gamma+\beta_0) + \sqrt{(c-1)^2 + 4c(\gamma+\beta_0)^2}]}{c(1-(\gamma+\beta_0)^2)}. \tag{6}$$

If $\gamma \geq 1 - \beta_0$, then no such optimal λ^* exists.

*(c) For any temperature $0 < \gamma < 1 - \beta_0$, at $\lambda = \lambda^*$, $\mathcal{F}_\infty^\gamma$ is **monotone decreasing** in $c \in (0, \infty)$.*

The expression for the asymptotic Bayes free energy $\mathcal{F}_\infty^\gamma$ is provided in the Supplementary Material. To summarize, first, in the spirit of empirical Bayes, there exists an optimal λ^* for the Gaussian prior which minimizes the asymptotic mean free energy. Under this setup, the choice of λ which maximizes the marginal likelihood for a particular realization of X, Y will converge almost surely to λ^* as $n, d \to \infty$. Similar to Nakkiran et al. (2020); Wu & Xu (2020), we find that model performance under marginal likelihood improves monotonically with input dimension when $\lambda = \lambda^*$ for a fixed amount of data. Indeed, for large n, d, $\mathbb{E}\mathcal{F}_n^\gamma \approx n\mathcal{F}_\infty^\gamma$ and $c \approx d/n$, so Theorem 1c implies that the expected Bayes free energy decreases (approximately) monotonically with the input dimension, provided n is fixed and the optimal regularizer λ^* is chosen.

Discussion of assumptions. The assumption that the kernel scales with λ is necessary using our techniques, as λ^* cannot be computed explicitly otherwise. This holds for the linear kernel, but most other choices of κ can be made to satisfy the conditions of Theorem 1 by taking $\kappa(x) \mapsto \eta^{-1}\kappa(\eta x)$, for appropriately chosen bandwidth $\eta \equiv \eta(\lambda)$. For example, for the quadratic kernel, this gives $k(x, x') = (\lambda^{-1/2} + \lambda^{1/2} x^\top x')^2$. Effectively, this causes the regularization parameter to scale non-linearly in the prior kernel. Even though this is required for our theory, we can empirically demonstrate this monotonicity also holds under the typical setup where k does not change with λ. In Figure 2, we plot the mean free energy for synthetic Gaussian datasets of increasing dimension at both optimal and fixed values of λ for the linear and Gaussian kernels. Since n is fixed, in line with Theorem 1c, the curves with optimally chosen λ decrease monotonically with input dimension, while the curves for fixed λ appear to increase when the dimension is large. Note, however, that the larger β for the Gaussian kernel induces a significant regularizing effect. A light CPE appears for the Gaussian kernel when λ is fixed, but does not seem to occur under λ^*.

While the assumption that $m = 0$ may appear too restrictive, in Appendix B, we show that m is necessarily small when the data is normalized and whitened. Consequently, under a zero-mean prior, the marginal likelihood behaves similarly to our assumed scenario. This translates well in practice: under a similar setup to Figure 2, the error curves corresponding to the linear and Gaussian kernels under the whitened `CIFAR10` benchmark dataset (Krizhevsky & Hinton, 2009) exhibiting the predicted monotone behavior (Figure 3).

Synthetic covariates. Since Theorem 1 implies that performance under the marginal likelihood can improve as covariates are added, it is natural to ask whether an improvement will be seen if

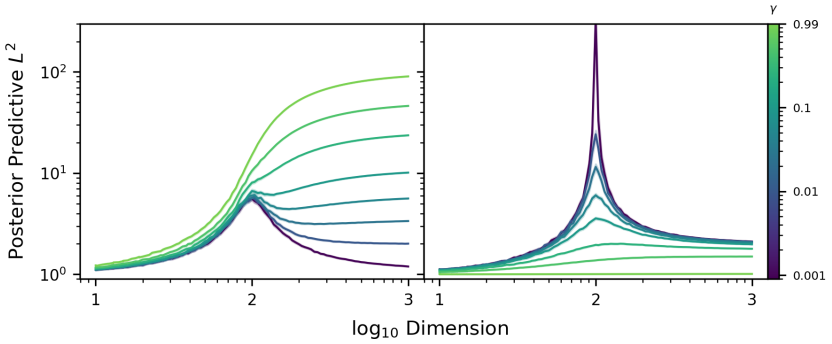

Figure 5: Posterior predictive L^2 loss error curves for **synthetic data** exhibiting perturbed / tempered double descent under the linear kernel with $\lambda = 0.01/\gamma$ (left), and $\lambda = \lambda^*$ (right).

the data is augmented with synthetic covariates. To test this, we considered the first 30 covariates of the whitened `CT Slices` dataset obtained from the UCI Machine Learning Repository (Graf et al., 2011), and we augmented them with synthetic (iid standard normal) covariates; the first 30 covariates repeated; and zeros (for more details, see Appendix A). While the first of these scenarios satisfies the conditions of Theorem 1, the second two do not, since the new data cannot be whitened such that its rows have unit covariance. Consequently, the behavior of the mean free energy reflects whether the assumptions of Theorem 1 are satisfied: only the data with Gaussian covariates exhibits the same monotone decay. From a practical point of view, a surprising conclusion is reached: after optimal regularization, performance under marginal likelihood can be further improved by concatenating Gaussian noise to the input.

5 Double Descent in Posterior Predictive Loss

In this section, we will demonstrate that, despite the connections between them, the marginal likelihood and posterior predictive loss can exhibit different qualitative behavior, with the posterior predictive losses potentially exhibiting a double descent phenomenon. Observe that the two forms of posterior predictive loss defined in (1) and (2) can both be expressed in the form

$$\mathcal{L} = c_1(\gamma)\underbrace{\mathbb{E}\|\bar{f}(\boldsymbol{x}) - \boldsymbol{y}\|^2}_{\text{MSE}} + c_2(\lambda,\gamma)\underbrace{\mathbb{E}\mathrm{tr}(\Sigma(\boldsymbol{x}))}_{\text{volume}} + c_3(\gamma).$$

The first term is the mean-squared error (MSE) of the predictor $\bar{f}$, and is a well-studied object in the literature. In particular, **the MSE can exhibit double descent**, or other types of multiple descent error curves depending on k, in both ridgeless (Holzmüller, 2020; Liang & Rakhlin, 2020) and general (Liu et al., 2021) settings. On the other hand, the volume term has the uniform bound $\mathbb{E}\mathrm{tr}(\Sigma(\boldsymbol{x})) \leq m\mathbb{E}k(x,x)$, so provided c_2 is sufficiently small, the volume term should have little qualitative effect. The following is immediate.

Proposition 1. *Assume that the MSE $\mathbb{E}\|\bar{f}(\boldsymbol{x}) - \boldsymbol{y}\|^2$ for Gaussian inputs $\boldsymbol{x}$ and labels $\boldsymbol{y}$ converges to an error curve $E(c)$ that exhibits double descent as $n \to \infty$ with $d \equiv d(n)$ satisfying $d(n)/n \to c \in (0,\infty)$. If there exists a function $\lambda(\gamma)$ such that $c_2(\lambda(\gamma),\gamma)/c_1(\gamma) \to 0$ as $\gamma \to 0^+$, then for any $\epsilon > 0$, there exists an error curve $\bar{E}(c)$ exhibiting double descent, a positive integer N, and $\gamma_0 > 0$ such that for any $0 < \gamma < \gamma_0$ and $n > N$, $|\mathcal{L}/c_1 - \bar{E}| < \epsilon$ at $d = d(n)$ and $\lambda = \lambda(\gamma)$.*

For **posterior predictive L^2 loss**, in the tempered posterior scenario where $\lambda = \mu/\gamma$, the MSE remains constant in γ, while $c_2/c_1 = \gamma/\mu$. Since the predictor $\bar{f}$ depends only on μ, the optimal γ in the tempered posterior scenario is realised as $\gamma \to 0^+$. In other words, under the posterior predictive L^2 loss, *the best prediction of uncertainty is none at all*. This highlights a trivial form of CPE for PPL2 losses, suggesting it may not be suitable as a UQ metric. Here we shall empirically examine the linear kernel case; similar experiments for more general kernels are conducted in the Supplementary Material. In Figure 5(left), we plot posterior predictive L^2 loss under the linear kernel on synthetic Gaussian data by varying γ while keeping μ fixed. We find that the error curve exhibits double descent when $\gamma < 2\mu$. The corresponding plot for the `CIFAR10` dataset is shown

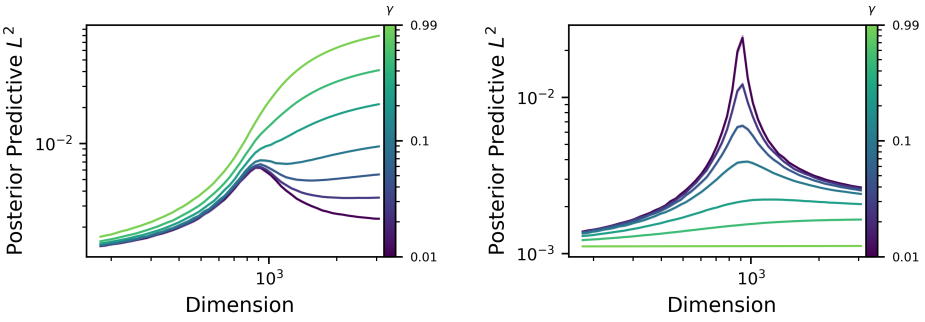

Figure 6: PPL2 loss under the linear kernel with $\lambda = 0.01/\gamma$ (left) and $\lambda = \lambda^*$ (right) on the CIFAR10 dataset; **curves for real data match Figure 5.**

in Figure 6(left), demonstrating that this behavior carries over to real data. Choosing $\lambda = \lambda^*$ (the optimal λ according to marginal likelihood) reveals a more typical set of regularized double descent curves; this is shown in Figure 5(right) for synthetic data and Figure 6(right) for the `CIFAR10` dataset. This is due to the monotone relationship between the volume term and λ, hence the error curve inherits its shape from the behaviour of λ^* (see Appendix A in the Supplementary Material).

In contrast, this phenomenon is not the case for **posterior predictive negative log-likelihood**. Indeed, letting $\lambda = \mu/\gamma$ and optimizing the expectation of (2) in γ, the optimal $\gamma^* = m^{-1}\mathbb{E}\|\bar{f}(\boldsymbol{x}) - \boldsymbol{y}\|^2$. The expected optimal PPNLL is therefore

$$-\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}\mathbb{E}_{f\sim\rho^{\gamma^*}}\log p(\boldsymbol{y}|f,\boldsymbol{x}) = \tfrac{1}{2}m[1+\log(2\pi\mathbb{E}\|\bar{f}(\boldsymbol{x})-\boldsymbol{y}\|^2)] + (2\mu)^{-1}\mathrm{tr}(\Sigma(\boldsymbol{x})). \quad (7)$$

Otherwise, the PPNLL displays similar behavior to PPL2, as the two are related linearly.

6 Conclusion

Motivated by understanding the uncertainty properties of prediction from GP models, we have applied random matrix theory arguments and conducted several experiments to study the error curves of three UQ metrics for GPs. Contrary to classical heuristics, model performance under marginal likelihood/Bayes free energy improves monotonically with input dimension under appropriate regularization (Theorem 1). However, Bayes free energy does not exhibit double descent. Instead, posterior predictive loss inherits a double descent curve from non-UQ settings when the variance in the posterior distribution is sufficiently small (Proposition 1). While our analysis was conducted under the assumption of a perfectly chosen prior mean, similar error curves appear to hold under small perturbations, which always holds for large whitened datasets. Although our contributions are predominantly theoretical, our results also have some noteworthy practical consequences:

- Tuning hyperparameters according to marginal likelihood is **essential** to ensuring good performance in higher dimensions, and **completely negates the curse of dimensionality**.
- When using L^2 losses as UQ metrics, care should be taken in view of the CPE. As such, **we do not recommend the use of this metric in lieu of other alternatives**.
- Our experiments suggest that further improvements beyond the optimisation of hyperparameters may be possible with the addition of synthetic covariates, although further investigation is needed before such a procedure can be universally recommended.

In light of the surprisingly complex behavior on display, the fine-scale behavior our results demonstrate, and a surprising absence of UQ metrics in the double descent literature, we encourage increasing adoption of random matrix techniques for studying UQ / Bayesian metrics in double descent contexts and beyond. There are numerous avenues available for future work, including the incorporation of more general kernels (e.g., using results from Fan & Wang (2020) to treat neural tangent kernels, which are commonly used as approximations for large-width neural networks).

REFERENCES

Ben Adlam and Jeffrey Pennington. The neural tangent kernel in high dimensions: Triple descent and a multi-scale theory of generalization. In *International Conference on Machine Learning*, pp. 74–84. PMLR, 2020.

Laurence Aitchison. A statistical theory of cold posteriors in deep neural networks. In *International Conference on Learning Representations*, 2020.

Ahmed Alaoui and Michael W Mahoney. Fast randomized kernel ridge regression with statistical guarantees. *Advances in Neural Information Processing Systems*, 28, 2015.

Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Mikhail Belkin, Daniel J. Hsu, and Partha Mitra. Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate. *Advances in neural information processing systems*, 31, 2018a.

Mikhail Belkin, Siyuan Ma, and Soumik Mandal. To understand deep learning we need to understand kernel learning. In *International Conference on Machine Learning*, pp. 541–549. PMLR, 2018b.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

Rajendra Bhatia. *Matrix analysis*, volume 169. Springer Science & Business Media, 2013.

Mickael Binois and Nathan Wycoff. A survey on high-dimensional Gaussian process modeling with application to Bayesian optimization. *arXiv preprint arXiv:2111.05040*, 2021.

Christopher M. Bishop and Nasser M. Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

Theophilos Cacoullos. On upper and lower bounds for the variance of a function of a random variable. *The Annals of Probability*, 10(3):799–809, 1982.

Wei-Cheng Chang, Chun-Liang Li, Yiming Yang, and Barnabás Póczos. Data-driven random fourier features using stein effect. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pp. 1497–1503, 2017.

Romain Couillet and Merouane Debbah. *Random matrix methods for wireless communications*. Cambridge University Press, 2011.

Stéphane d'Ascoli, Levent Sagun, and Giulio Biroli. Triple descent and the two kinds of overfitting: Where & why do they appear? *Advances in Neural Information Processing Systems*, 33:3058–3069, 2020.

Filip De Roos, Alexandra Gessner, and Philipp Hennig. High-dimensional Gaussian process inference with derivatives. In *International Conference on Machine Learning*, pp. 2535–2545. PMLR, 2021.

Michal Derezinski, Rajiv Khanna, and Michael W. Mahoney. Improved guarantees and a multiple-descent curve for Column Subset Selection and the Nystrom method. In *Annual Advances in Neural Information Processing Systems 33: Proceedings of the 2020 Conference*, pp. 4953–4964, 2020a.

Michal Derezinski, Feynman T. Liang, and Michael W. Mahoney. Exact expressions for double descent and implicit regularization via surrogate random design. *Advances in neural information processing systems*, 33:5152–5164, 2020b.

Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279, 2018.

Alan Edelman and N. Raj Rao. Random Matrix Theory. *Acta numerica*, 14:233–297, 2005.

Noureddine El Karoui. The spectrum of kernel random matrices. *The Annals of Statistics*, 38(1):1 – 50, 2010.

Andreas Engel and Christian P. L. Van den Broeck. *Statistical mechanics of learning*. Cambridge University Press, New York, NY, USA, 2001.

Zhou Fan and Zhichao Wang. Spectra of the conjugate kernel and neural tangent kernel for linear-width neural networks. *Advances in Neural Information Processing Systems*, 33:7710–7721, 2020.

Edwin Fong and Chris C. Holmes. On the marginal likelihood and cross-validation. *Biometrika*, 107(2):489–496, 2020.

Vincent Fortuin, Adrià Garriga-Alonso, Florian Wenzel, Gunnar Rätsch, Richard Turner, Mark van der Wilk, and Laurence Aitchison. Bayesian neural network priors revisited. *arXiv preprint arXiv:2102.06571*, 2021.

Peter I. Frazier. A tutorial on Bayesian optimization. *arXiv preprint arXiv:1807.02811*, 2018.

Spencer Frei, Niladri S. Chatterji, and Peter L. Bartlett. Benign overfitting without linearity: Neural network classifiers trained by gradient descent for noisy linear data. *arXiv preprint arXiv:2202.05928*, 2022.

Pascal Germain, Francis Bach, Alexandre Lacoste, and Simon Lacoste-Julien. PAC-Bayesian theory meets Bayesian inference. *arXiv preprint arXiv:1605.08636*, 2016.

Ethan Goan and Clinton Fookes. Bayesian neural networks: An introduction and survey. In *Case Studies in Applied Bayesian Data Science*, pp. 45–87. Springer, 2020.

Franz Graf, Hans-Peter Kriegel, Matthias Schubert, Sebastian Pölsterl, and Alexander Cavallaro. 2D image registration in CT images using radial image descriptors. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 607–614. Springer, 2011.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986, 2022.

Ali Hebbal, Loic Brevault, Mathieu Balesdent, El-Ghazali Talbi, and Nouredine Melab. Bayesian optimization using deep Gaussian processes. *arXiv preprint arXiv:1905.03350*, 2019.

David Holzmüller. On the universality of the double descent peak in ridgeless regression. In *International Conference on Learning Representations*, 2020.

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

Konstantin Krivoruchko and Alexander Gribov. Evaluation of empirical Bayesian kriging. *Spatial Statistics*, 32:100368, 2019.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. 2009.

Loic Le Gratiet and Josselin Garnier. Asymptotic analysis of the learning curve for Gaussian process regression. *Machine learning*, 98(3):407–433, 2015.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Armin Lederer, Jonas Umlauft, and Sandra Hirche. Posterior variance analysis of Gaussian processes with application to average learning curves. *arXiv preprint arXiv:1906.01404*, 2019.

Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel "ridgeless" regression can generalize. *The Annals of Statistics*, 48(3):1329–1347, 2020.

Zhenyu Liao, Romain Couillet, and Michael W. Mahoney. A random matrix analysis of random Fourier features: beyond the Gaussian kernel, a precise phase transition, and the corresponding double descent. *Advances in Neural Information Processing Systems*, 33:13939–13950, 2020.

Fanghui Liu, Zhenyu Liao, and Johan Suykens. Kernel regression in high dimensions: Refined analysis beyond double descent. In *International Conference on Artificial Intelligence and Statistics*, pp. 649–657. PMLR, 2021.

Marco Loog, Tom Viering, Alexander Mey, Jesse H. Krijthe, and David M.J. Tax. A brief prehistory of double descent. *Proceedings of the National Academy of Sciences*, 117(20):10625–10626, 2020.

Charles H. Martin and Michael W. Mahoney. Rethinking generalization requires revisiting old ideas: statistical mechanics approaches and complex learning behavior. Technical Report Preprint: arXiv:1710.09553, 2017.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766, 2022.

Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, Arthur Gretton, and Bernhard Schölkopf. Kernel mean estimation and stein effect. In *International Conference on Machine Learning*, pp. 10–18. PMLR, 2014.

Vidya Muthukumar, Kailas Vodrahalli, Vignesh Subramanian, and Anant Sahai. Harmless interpolation of noisy data in regression. *IEEE Journal on Selected Areas in Information Theory*, 1(1): 67–83, 2020.

Preetum Nakkiran, Prayaag Venkat, Sham M. Kakade, and Tengyu Ma. Optimal regularization can mitigate double descent. In *International Conference on Learning Representations*, 2020.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003, 2021.

Lorenzo Noci, Kevin Roth, Gregor Bachmann, Sebastian Nowozin, and Thomas Hofmann. Disentangling the roles of curation, data-augmentation and the prior in the cold posterior effect. *Advances in Neural Information Processing Systems*, 34, 2021.

Manfred Opper and Francesco Vivarelli. General bounds on Bayes errors for regression with Gaussian processes. *Advances in Neural Information Processing Systems*, 11, 1998.

Debashis Paul and Alexander Aue. Random matrix theory in statistics: A review. *Journal of Statistical Planning and Inference*, 150:1–29, 2014.

Constantine Pozrikidis. *An introduction to grids, graphs, and networks*. Oxford University Press, 2014.

Stephen Roberts, Michael Osborne, Mark Ebden, Steven Reece, Neale Gibson, and Suzanne Aigrain. Gaussian processes for time-series modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1984):20110550, 2013.

Peter Sollich. Learning curves for Gaussian processes. *Advances in neural information processing systems*, 11, 1998.

Peter Sollich and Anason Halees. Learning curves for Gaussian process regression: approximations and bounds. *Neural Computation*, 14(6):1393–1428, 2002.

Stefano Spigler, Mario Geiger, and Matthieu Wyart. Asymptotic learning curves of kernel methods: empirical data versus teacher–student paradigm. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(12):124001, 2020.

William E Strawderman. On charles stein's contributions to (in) admissibility. *The Annals of Statistics*, 49(4):1823–1835, 2021.

Alexander Tsigler and Peter L. Bartlett. Benign overfitting in ridge regression. *arXiv preprint arXiv:2009.14286*, 2020.

Tom Viering and Marco Loog. The shape of learning curves: a review. *arXiv preprint arXiv:2103.10948*, 2021.

Ulrike von Luxburg and Olivier Bousquet. Distance-based classification with lipschitz functions. *J. Mach. Learn. Res.*, 5(Jun):669–695, 2004.

Ke Wang, Vidya Muthukumar, and Christos Thrampoulidis. Benign overfitting in multiclass classification: All roads lead to interpolation. *Advances in Neural Information Processing Systems*, 34, 2021.

Florian Wenzel, Kevin Roth, Bastiaan S. Veeling, Jakub Świątkowski, Linh Tran, Stephan Mandt, Jasper Snoek, Tim Salimans, Rodolphe Jenatton, and Sebastian Nowozin. How good is the Bayes posterior in deep neural networks really? *arXiv preprint arXiv:2002.02405*, 2020.

Christopher K. I. Williams and David Barber. Bayesian classification with Gaussian processes. *IEEE Transactions on pattern analysis and machine intelligence*, 20(12):1342–1351, 1998.

Christopher K. I. Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

Christopher K. I. Williams and Francesco Vivarelli. Upper and lower bounds on the learning curve for Gaussian processes. *Machine Learning*, 40(1):77–102, 2000.

Andrew G. Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. *Advances in neural information processing systems*, 33:4697–4708, 2020.

Denny Wu and Ji Xu. On the optimal weighted ℓ_2 regularization in overparameterized linear regression. *Advances in Neural Information Processing Systems*, 33:10112–10123, 2020.

Jiaxin Zhang. Modern Monte Carlo methods for efficient uncertainty quantification and propagation: a survey. *Wiley Interdisciplinary Reviews: Computational Statistics*, 13(5):e1539, 2021.

Tong Zhang. Learning bounds for kernel regression using effective data dimensionality. *Neural Computation*, 17(9):2077–2098, 2005.

Monotonicity and Double Descent in Uncertainty Quantification with Gaussian Processes

SUPPLEMENTARY DOCUMENT

A Additional Empirical Results

In this section, we consider other factors not covered by our analysis in the main body of the paper. Full experimental details are given in Appendix G.

`CT Slices` dataset. To demonstrate our procedure for working with real data, we first consider the `CT Slices` dataset obtained from the UCI Machine Learning Repository (Graf et al., 2011), comprised of $n = 53500$ images $X_1, \dots, X_n \in \mathbb{R}^d$ with $d = 385$ features, and corresponding scalar-valued labels $Y_1, \dots, Y_n \in \mathbb{R}$. This dataset is also used in Figure 4. The data was preprocessed in the following way: first, 17 features were observed to be linearly dependent on the others, and were removed to reveal $d = 368$ features. The sample mean μ_X and sample covariance matrix Σ_X of $X_1, \dots, X_n$ were computed, and the input normalized by $X_i \mapsto \Sigma_X^{-1/2}(X_i - \mu_X)$. The labels were similarly normalized as $Y_i \mapsto (Y_i - \mu_Y)/\sigma_Y$, where μ_Y and σ_Y are the sample mean and standard deviation of the labels, respectively. Under this preprocessing, X and Y are assumed to satisfy the conditions of Theorem 1.

Figure 7 examines the mean Bayes free energy for the linear and Gaussian kernels, under the optimal choice of λ^* from Theorem 1. This figure should be compared to the synthetic data examples shown in Figure 2 (upper left and bottom left). Similarly, Figure 8 is the corresponding version of Figure 5. Notably, the characteristic behavior of all four plots is still prominent in the real data example.

Image classification datasets We conducted parallel experiments on two larger benchmark datasets that are ubiquitous in the literature — `MNIST` LeCun et al. (1998) and `CIFAR10` Krizhevsky & Hinton (2009). To this end, the `MNIST` and `CIFAR10` datasets were preprocessed in the same manner as the `CT Slices` dataset. Both datasets correspond to classification problems with 10 class labels, however, for our purposes we consider the analogous regression problems over the class labels.

The `MNIST` training set is comprised of $60,000$ different 28×28 grayscale images of handwritten digits from 0-9. After preprocessing, $d = 706$ of the 768 features were retained, and $n = 175$ images were randomly sampled for use as the dataset. The mean free energy curves under the linear and Gaussian kernel under the optimal $\lambda = \lambda^*$, as well as the PPL2 curves for the optimal λ and fixed μ are shown in Figures 9 and 10, respectively. Similarly, the `CIFAR10` training dataset contains $50,000$ different 32×32 color images, each with 3 channels. This corresponds to 3072 features, of which $d = 3003$ were retained after preprocessing, and $n = 900$ images were randomly sampled as the for use as the dataset. Analogous images are presented in Figures 3 and 6.

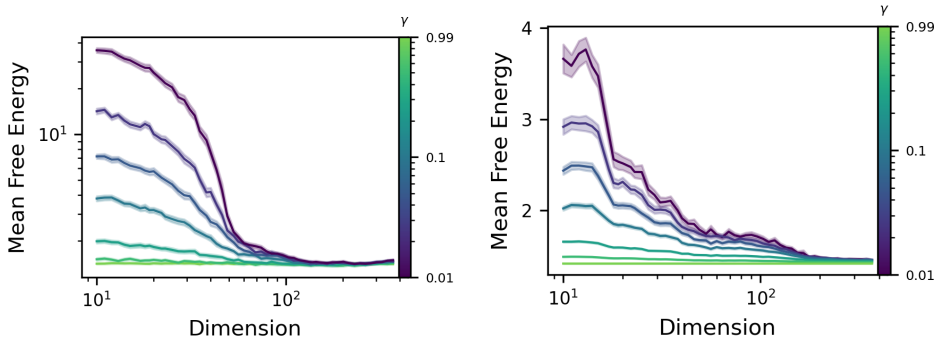

Figure 7: Error curves for mean Bayes free energy under the CT Slices dataset; linear (left) and Gaussian (right) kernels; $\lambda = \lambda^*$

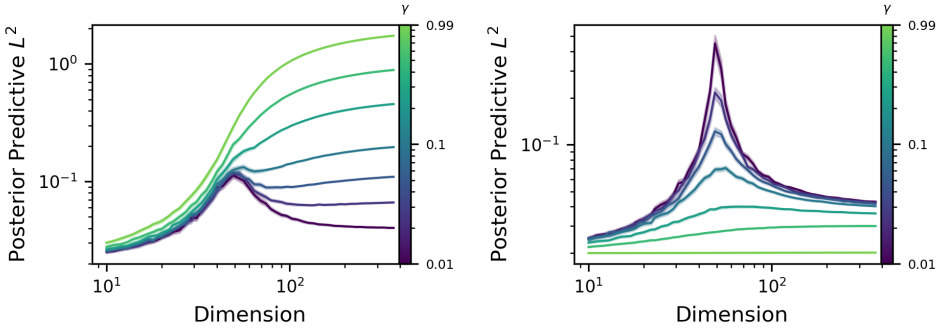

Figure 8: PPL2 loss under the linear kernel with $\lambda = 0.01/\gamma$ (left) and $\lambda = \lambda^*$ (right) on the CT Slices dataset.

It may seem surprising that the behavior of these models is so close to those of well-specified models, since there is no a priori reason to assume the mean of the data-generating process is zero. However, in Appendix B we demonstrate that this is merely a consequence of normalization of the response variables, and that such normalization forces tight control over the gradients of the mean function under expectation.

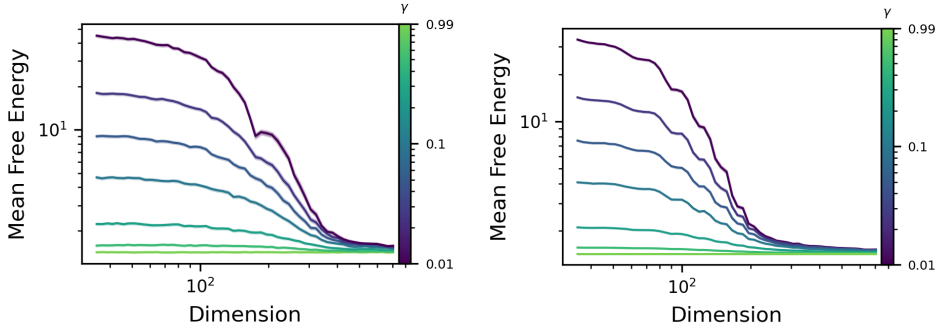

Figure 9: Error curves for mean Bayes free energy under the MNIST dataset; linear (left) and Gaussian (right) kernels; $\lambda = \lambda^*$

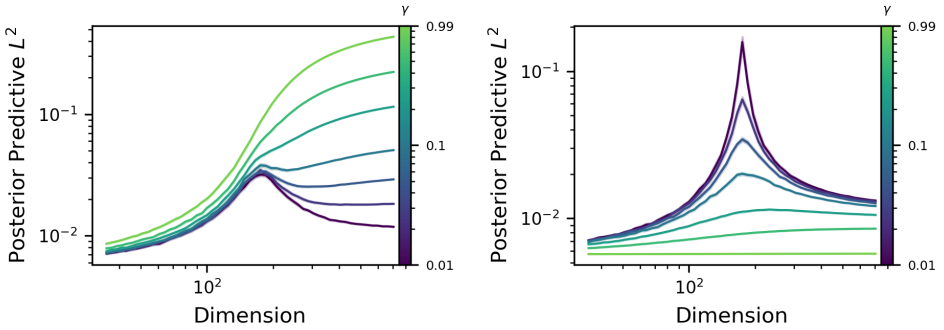

Figure 10: PPL2 loss under the linear kernel with $\lambda = 0.01/\gamma$ (left) and $\lambda = \lambda^*$ (right) on the MNIST dataset.

Synthetic covariates. From Theorem 1, one can conclude that performance under the marginal likelihood can increase as covariates are added. This begs the question: if the data is augmented with synthetic covariates, will this still result in a higher marginal likelihood? We have considered adding three different forms of synthetic covariates to the first 30 covariates of the whitened `CT Slices` dataset:

(i) **Gaussian white noise**: each X_{ij} for $j > 30$ is drawn as an iid standard normal random variable;

(ii) **Copied data**: the first 30 covariates are repeated, that is, for $j > 30$, each $X_{ij} = X_{i,(j-1) \bmod 30+1}$, where mod denotes the modulus operator; and

(iii) **Padded data**: each $X_{ij} = 0$ for $j > 30$.

While case (i) satisfies the conditions of Theorem 1, cases (ii) and (iii) do not, as neither case can be whitened such that the rows of X have unit covariance. In Figure 4, we repeat the experiment in the top left of Figure 2 using these augmented datasets. The behavior of the mean Bayes free energy reflects whether the assumptions of Theorem 1 are satisfied: while case (i) exhibits the same monotone decay, cases (ii) and (iii) do not.

Monotonicity in posterior predictive metrics. In these experiments, we consider posterior predictive metrics for synthetic data under optimal parameter choices. First, in Figure 11, the posterior predictive L^2 loss is optimized in λ, revealing a monotone decay in the dimension, analogous to Nakkiran et al. (2020); Wu & Xu (2020). In Figure 12, we plot error curves for the optimally tempered PPNLL metric (7) under the linear kernel, revealing a monotonically increasing curve with input dimension when μ is fixed, and highlighting the need for appropriate regularization. If PPNLL is optimized in both γ and μ simultaneously, the error curve becomes flat.

Prior misspecification. In our analysis, we have considered a practically optimal scenario where the prior is centered on the mean function of the labels (in other words, our prior concentrates on the correct solution). For more complex setups, where the prior is implicit and data-dependent, this may be possible, but is unlikely in general. For example, if the prior dictates *a priori* knowledge, then a perfectly specified prior implies the underlying generative model for the labels is known in advance. Here, we assume that the mean function of the labels is nonzero, emulating a more realistic scenario. We restrict ourselves to the linear setting here, and we consider $Y_i = \theta_0 X_i + \epsilon_i$, ensuring that the correct mean function lies in the RKHS of the kernel. Figure 13 illustrates the effect on Bayes free energy (with optimal λ^*). From left to right, small $\theta_0 = d^{-1/2}\mathbf{1}$, large $\theta_0 = nd^{-1/2}\mathbf{1}$, and growing $\theta_0 = \mathbf{1}$ perturbations are considered. For small values of the perturbation, the monotonicity of the error curve is not affected in a meaningful way. While the zero-mean assumption may seem restrictive, we demonstrate that this scenario will always hold asymptotically, provided the data is normalized and whitened (see Appendix B). For larger perturbations, however, we see a horizontal "double-ascent" (or reverse double descent) error curve. A growing perturbation also results in a double-ascent curve, but with increasing Bayes free energy once the input dimension is sufficiently large.

Regularity of the kernel. The regularity of the kernel plays a key role in the regularity, and consequently, the quality of the predictor. In particular, less regular predictors tend to revert to the prior more quickly away from the training data. The Matérn kernel family is noteworthy for its capacity to adjust the regularity of predictors through the parameter ν, whereby realizations of a Gaussian process with Matérn covariance are at most $\lceil \nu \rceil$-times differentiable (Williams & Rasmussen, 2006,

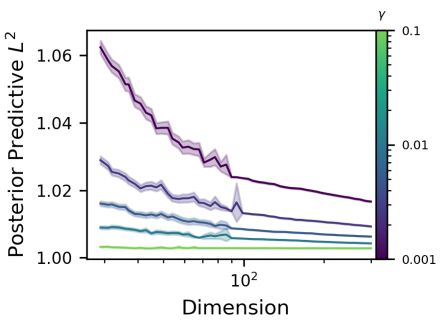

Figure 11: PPL2 optimized in λ; varying γ.

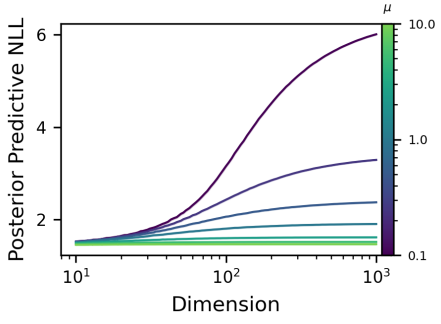

Figure 12: PPNLL optimized in γ with $\lambda = \mu/\gamma$; varying μ.

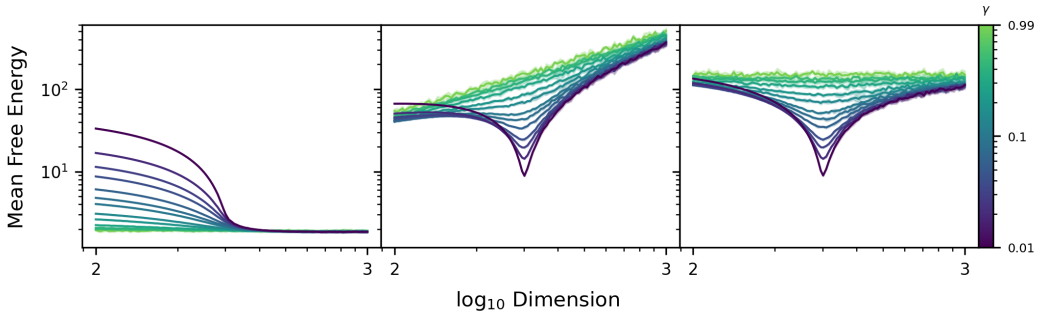

Figure 13: Optimal mean Bayes free energy with low (left), increasing (center) and high (right) levels of prior misspecification under the linear kernel.

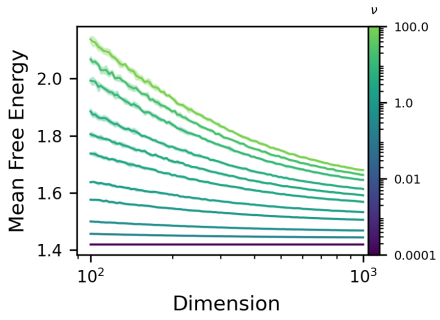

Figure 14: Effect of varying the regularity parameter ν in the Matérn kernel

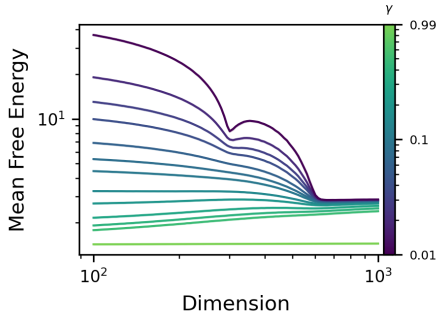

Figure 15: Effect of irregular spectra in X under the linear kernel

pg. 85). In Figure 14, we plot the Bayes free energy for fixed n = 300 and $\gamma = 0.01$ with optimal λ^* over input dimensions $d \in [100, 1000]$ and $\nu \in [0.5, 100]$. As ν decreases, the curves become flatter, suggesting the effect of dimension is reduced.

Ill-conditioned data. Our theoretical analysis considers only the case where the data has been whitened, that is, where each row of X has unit covariance. It is known that more interesting behavior can occur depending on the spectrum of eigenvalues of the covariance matrix, including multiple descent (Nakkiran et al., 2020; Hastie et al., 2022), and this appears to be robust to other volume-based objectives (Derezinski et al., 2020a). In Figure 15, we consider an isotropic ill-conditioned covariance matrix $\mathrm{Cov}(X_i) = \Sigma$ where $\Sigma = \mathrm{diag}((10)_{i=1}^{d/2}, (1/10)_{i=1}^{d/2})$. Under the linear kernel, for fixed λ, the error curve is similar to the isotropic setting. However, at $\lambda = \lambda^*$, we find that the mean Bayes free energy can exhibit non-monotonic behavior at low temperatures.

Scaling dimension nonlinearly with data. An interesting consequence of the monotonic error curve in the Bayes free energy is that the inclusion of additional data may be harmful if the input dimension is increased at a slower rate $d = \mathcal{O}(n^\xi)$ for $\xi < 1$ (or beneficial if $\xi > 1$). This effect is illustrated in Figure 16, where the normalized Bayes free energy $n^{-1}\mathcal{F}_n^\gamma$ is plotted for the linear and Gaussian kernels at the optimal λ^* over $n \in [300, 3000]$ with $d = 2^{10(1-\xi)}n^\xi$.

Effect of noise distribution Each experiment has also assumed that the labels are standard normal. If this is not the case, but the labels are still assumed to be iid, have zero mean and are uncorrelated

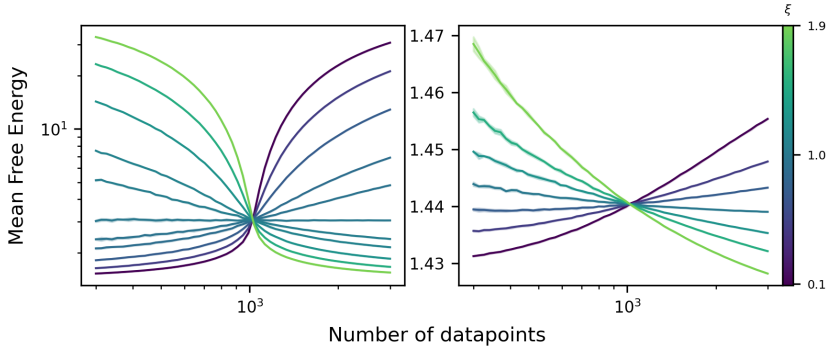

Figure 16: Error curves for mean Bayes free energy $n^{-1}\mathcal{F}_n^\gamma$ under linear (left) and Gaussian (right) kernels and $\lambda = \lambda^*$, for dimension scaling with data as $\mathcal{O}(n^\xi)$

with the inputs (correctly specified prior), then the expected mean Bayes free energy satisfies

$$n^{-1}\mathbb{E}\mathcal{F}_n^\gamma = \frac{\lambda}{2n}\sum_{i,j=1}^{n}\mathbb{E}[Y_iY_jQ_{ij}] + \frac{1}{2n}\mathbb{E}\log\det(K_X+\lambda\gamma I) - \frac{1}{2}\log\left(\frac{\lambda}{2\pi}\right),$$
$$= \frac{\lambda}{2n}\sigma^2\mathbb{E}\mathrm{tr}(Q) + \frac{1}{2n}\mathbb{E}\log\det(K_X+\lambda\gamma I) - \frac{1}{2}\log\left(\frac{\lambda}{2\pi}\right),$$

where $Q = (K_X + \lambda\gamma I)^{-1}$ and $\sigma^2 = \mathbb{E}[Y_i^2]$. Therefore, only the variance in the labels contributes to $n^{-1}\mathbb{E}\mathcal{F}_n^\gamma$ (other features of the distribution of the noise contribute to the higher order moments of $\mathcal{F}_n^\gamma$). In Figure 17, we examine the effect that different variances in the label noise have on the mean Bayes free energy. Normally distributed Y_i were considered, with variances ranging from 0.1 to 10.

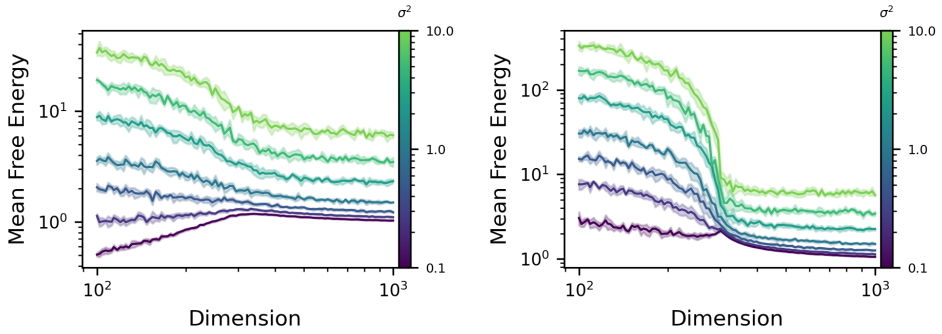

Figure 17: Error curves for mean Bayes free energy under linear kernel with $\lambda = \lambda^*$, $\gamma = 0.1$ (left) / $\gamma = 0.01$ (right) and different variances in the label data.

Posterior predictive loss with Gaussian kernel Figures 18 and 19 examine the effect of varying γ on the posterior predictive L^2 loss varying over d, under the Gaussian kernel. These figures should be contrasted with the linear kernel case presented as Figure 5 (left and right, respectively). Note that the significant regularizing effect when $\beta > 0$ prohibits the double descent behavior found in the linear kernel case.

Visualizing λ^* Figures 20 and 21 plot the values of λ^* versus c over different values of γ, for the linear and Gaussian kernels, respectively. Once again, the sharp trough formed at $d = n$ in Figure 21 is significantly dampened by the regularizing effect of $\beta > 0$.

Real data without whitening. To examine the effect that whitening has on the error curves, we reconsider the experiments producing Figures 3 (left; MNIST) and 9 (left; CIFAR10) where X and Y are only *normalised*, that is, we subtract the sample means and divide by the sample deviation. The results are reported in Figure 22. As expected from Couillet & Debbah (2011) and the results of Figure 15, the curves resemble their whitened counterparts with some spurious "bumps".

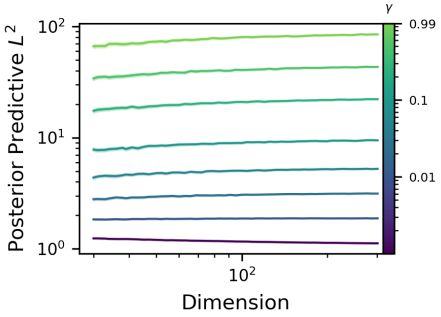

Figure 18: Posterior predictive L^2 loss under the Gaussian kernel with $\lambda = 0.01/\gamma$

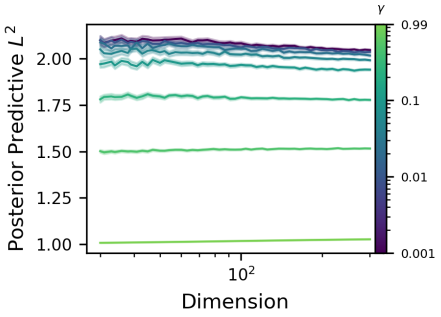

Figure 19: Posterior predictive L^2 loss under the Gaussian kernel with $\lambda = \lambda^*$

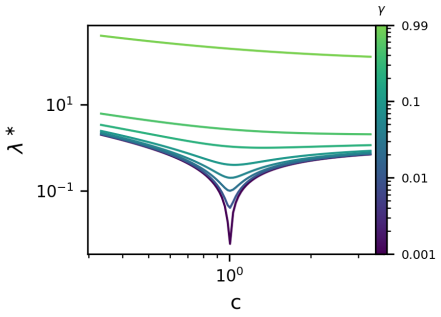

Figure 20: Values of λ^* versus c varying γ for the linear kernel

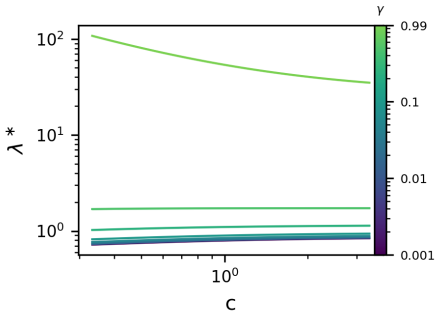

Figure 21: Values of λ^* versus c varying γ for the Gaussian kernel

B Normalized Data implies Small Prior Misspecification

In Figure 13, we explored the effect of changing the mean of the data-generating process from that of the prior. It was found that provided the mean of the data-generating process did not differ too significantly from that of the prior, the monotonicity of the error curves in Bayes free energy appeared unaffected. Here we show that when the data is normalized and whitened, and a zero-mean prior is chosen, the mean of the data-generating process will *never* differ too significantly from the prior.

As above, assume that the labels satisfy $Y_i = f(X_i) + \epsilon_i$ for some $f : \mathbb{R}^d \to \mathbb{R}$ and zero-mean iid ϵ_i. Now, we also assume that Y has been normalized so that $\text{Var}(Y) = 1$. Similarly, we assume that X has been normalized and whitened, so that it has zero mean and unit covariance. For simplicity of

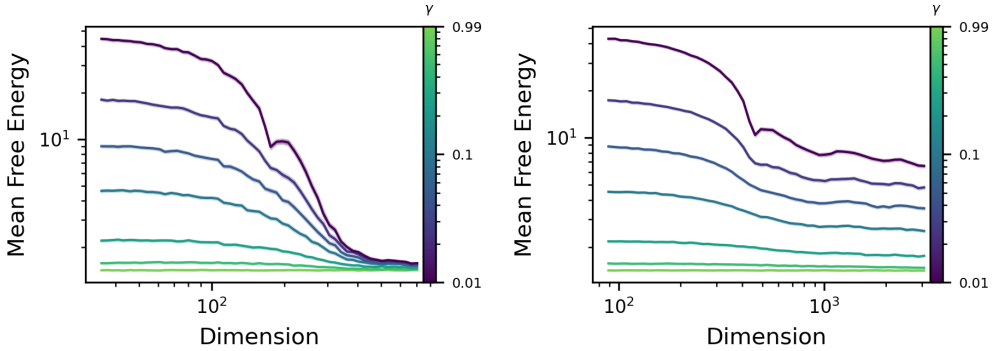

Figure 22: Error curves for mean Bayes free energy for the MNIST (left) and CIFAR10 (right) datasets, with linear kernels; $\lambda = \lambda^*$.

argument, assume further that X_i are normal, that is, $X_i \overset{\text{iid}}{\sim} Z \sim \mathcal{N}(0, I)$. In the linear case where $f(x) = \theta \cdot x$, since

$$1 = \text{Var}(Y) \geq \text{Var} f(Z) = \|\theta\|^2,$$

this implies that the magnitude of the components of θ are bounded on average by $d^{-1/2}$. This is the scenario seen in Figure 13(left). Indeed, the scenarios in Figure 13(center) and 13(right), which exhibit different error curves, satisfy $\text{Var}(Y) = n$ and $\text{Var}(Y) = d$ respectively, both of which are considerably larger than 1.

The same principle holds for more general f. By a reverse Gaussian Poincaré inequality (Cacoullos, 1982, Proposition 3.5),

$$1 = \text{Var}(Y) \geq \text{Var} f(Z) \geq \frac{1}{d} \left(\sum_{i=1}^{d} \mathbb{E} \partial_i f(Z) \right)^2,$$

where ∂_i denotes the i-th partial derivative. Therefore, the average coordinate-wise gradient of f, $\mathbb{E} \partial_I f(X)$ (where I is uniform over $\{1, \ldots, d\}$), is bounded above and below by

$$-\sqrt{\frac{1}{d}} \leq \mathbb{E} \partial_I f(X) = \frac{1}{d} \sum_{i=1}^{d} \mathbb{E} \partial_i f(X) \leq \sqrt{\frac{1}{d}}.$$

C Digamma Function

Before treating the random matrix theory, we will need some auxiliary results concerning the *digamma function*. Let $\Gamma(z)$ be the Gamma function, defined for $z > 0$ by $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} \mathrm{d}t$. The *digamma function* $\psi(z)$ is the derivative of the logarithm of the Gamma function, that is $\psi(z) = \frac{\mathrm{d}}{\mathrm{d}z} \log \Gamma(z)$. The digamma function satisfies the following properties:

- $\psi(z+1) = \psi(z) + z^{-1}$ for any $z > 0$;
- as $z \to \infty$, $\psi(z) / \log z \to 1$;
- letting $\gamma_{\text{EM}} = \psi(1)$ denote the Euler-Mascheroni constant,

$$\psi(z+1) = -\gamma_{\text{EM}} + \int_0^1 \left(\frac{1 - t^z}{1 - t} \right) \mathrm{d}t.$$

The digamma function behaves well under summation. In particular, we have the following lemma.

Lemma 1. *For any positive integer n and any real number $z > -1$,*

$$\sum_{i=1}^{n} \psi(z+i) = (n+z)\psi(n+z) - z\psi(z) - n.$$

Proof. From the integral representation for the digamma function:

$$\psi(z) = -\gamma_{\text{EM}} + \int_0^1 \left(\frac{1 - t^{z-1}}{1 - t} \right) \mathrm{d}t,$$

since $\sum_{i=1}^{n} t^{z+i-1} = t^z \sum_{i=0}^{n-1} t^i = t^z \frac{1-t^n}{1-t}$ for $0 \leq t < 1$,

$$\sum_{i=1}^{n} \psi(z+i) = -n\gamma_{\text{EM}} + \int_0^1 \frac{n(1-t) - t^z(1-t^n)}{(1-t)^2} \mathrm{d}t.$$

Focusing on the integral term, note that by letting $f(t) = n(1-t) - t^z(1-t^n)$ and $g(t) = (1-t)^{-1}$, since $g'(t) = (1-t)^{-2}$,

$$\begin{aligned} \int_0^1 \frac{n(1-t) - t^z(1-t^n)}{(1-t)^2} \mathrm{d}t &= \int_0^1 f(t) g'(t) \mathrm{d}t \\ &= \lim_{t \to 1^-} f(t)g(t) - f(0)g(0) - \int_0^1 f'(t) g(t) \mathrm{d}t. \end{aligned}$$

Since $\lim_{t\to 1^-}(1-t^n)/(1-t) = n$, $\lim_{t\to 1^-} f(t)g(t) = 0$, and so

$$\begin{aligned}
\int_0^1 \frac{n(1-t) - t^z(1-t^n)}{(1-t)^2}\mathrm{d}t &= -n - \int_0^1 \frac{-n - zt^{z-1} + (n+z)t^{n+z-1}}{1-t}\mathrm{d}t \\
&= -n - \int_0^1 \frac{(n+z)(t^{n+z-1}-1) - z(t^{z-1}-1)}{1-t}\mathrm{d}t \\
&= -n + (n+z)\int_0^1 \frac{1-t^{n+z-1}}{1-t}\mathrm{d}t - z\int_0^1 \frac{1-t^{z-1}}{1-t}\mathrm{d}t \\
&= -n + (n+z)\left[\psi(n+z) + \gamma_{\text{EM}}\right] - z\left[\psi(z) + \gamma_{\text{EM}}\right] \\
&= -n + n\gamma_{\text{EM}} + (n+z)\psi(n+z) - z\psi(z).
\end{aligned}$$

The result immediately follows □

Using this lemma, we can obtain an explicit expression for the sum of digamma functions with increment $\frac{1}{2}$. This will be particularly useful for computing determinants of Wishart matrices.

Lemma 2. *For any positive integers n and d with $n > d$,*

$$\begin{aligned}
\sum_{i=1}^{d}\psi\left(\frac{n-i+1}{2}\right) = \frac{n}{2}\psi\left(\frac{n}{2}\right) &- \left(\frac{n-d}{2}\right)\psi\left(\frac{n-d}{2}\right) - d \\
&+ \left(\frac{n-1}{2}\right)\psi\left(\frac{n-1}{2}\right) \\
&- \left(\frac{n-d-1}{2}\right)\psi\left(\frac{n-d-1}{2}\right).
\end{aligned}$$

Proof. First, note that

$$\sum_{i=1}^{d}\psi\left(\frac{n-i+1}{2}\right) = \sum_{i=1}^{d}\psi\left(\frac{n-d+i}{2}\right).$$

We consider the cases where d is even and odd separately. When d is even,

$$\begin{aligned}
\sum_{i=1}^{d}\psi\left(\frac{n-d}{2}+\frac{i}{2}\right) &= \sum_{i=1}^{d/2}\psi\left(\frac{n-d}{2}+i\right) + \psi\left(\frac{n-d}{2}+i-\frac{1}{2}\right) \\
&= \frac{n}{2}\psi\left(\frac{n}{2}\right) - \left(\frac{n-d}{2}\right)\psi\left(\frac{n-d}{2}\right) - d \\
&\quad + \left(\frac{n-1}{2}\right)\psi\left(\frac{n-1}{2}\right) \\
&\quad - \left(\frac{n-d-1}{2}\right)\psi\left(\frac{n-d-1}{2}\right).
\end{aligned}$$

Now assume that d is odd. Then

$$\begin{aligned}
\sum_{i=1}^{d}\psi\left(\frac{n-d}{2}+\frac{i}{2}\right) &= \psi\left(\frac{n}{2}\right) + \sum_{i=1}^{d-1}\psi\left(\frac{(n-1)-(d-1)}{2}+\frac{i}{2}\right) \\
&= \psi\left(\frac{n}{2}\right) + \frac{n-1}{2}\psi\left(\frac{n-1}{2}\right) - \left(\frac{n-d}{2}\right)\psi\left(\frac{n-d}{2}\right) - d + 1 \\
&\quad + \left(\frac{n}{2}-1\right)\psi\left(\frac{n}{2}-1\right) - \left(\frac{n-d-1}{2}\right)\psi\left(\frac{n-d-1}{2}\right).
\end{aligned}$$

But now, since $z\psi(z+1) = z\psi(z) + 1$, $(\frac{n}{2}-1)\psi(\frac{n}{2}-1) = (\frac{n}{2}-1)\psi(\frac{n}{2}) - 1$, and so

$$\psi\left(\frac{n}{2}\right) + \left(\frac{n}{2}-1\right)\psi\left(\frac{n}{2}-1\right) = \frac{n}{2}\psi\left(\frac{n}{2}\right) - 1.$$

The result now follows.

□

D MARCHENKO-PASTUR THEORY

In this section, we prove several lemmas concerning limiting traces and log-determinants of Wishart matrices that will prove foundational for proving our main results. The fundamental theorem in this section is the Marchenko-Pastur Theorem, which describes the limiting spectral distribution of Wishart matrices. The following can be obtained from pg. 51 of Couillet & Debbah (2011).

Theorem 2 (Marchenko-Pastur Theorem)**.** *For each $n = 1, 2, \ldots$, let $X_n \in \mathbb{R}^{n\times d}$ be a matrix of iid random variables with zero mean and unit variance. If $n, d \to \infty$ with $d/n \to c \in (0, \infty)$, then for every $z \in \mathbb{C}\backslash\{0\}$,*

$$d^{-1}\mathbb{E}\mathrm{tr}((n^{-1}X_n^\top X_n - zI)^{-1}) \to m(z) \coloneqq \frac{1 - c - z - \sqrt{(z-c-1)^2 - 4c}}{2cz},$$

noting that $m(z)$ satisfies $m = 1/(1 - c - z - czm)$.

For the remainder of this section, we assume the conditions of Theorem 2, that is, for each $n = 1, 2, \ldots$, we let $X_n \in \mathbb{R}^{n\times d}$ be a matrix of iid random variables with zero mean and unit variance.

Lemma 3 (Trace of Inverse Matrix)**.** *Let $n, d \to \infty$ with $d/n \to c \in (0, 1]$ and assume that μ_n is a sequence of real numbers such that $\mu_n \to \mu \in (0, \infty)$ as $n \to \infty$. Then*

$$n^{-1}\mathbb{E}\mathrm{tr}((d^{-1}X_n^\top X_n + \mu_n I)^{-1}) \to T(\mu, c) \coloneqq \frac{c - 1 - c\mu + \sqrt{(c\mu + c + 1)^2 - 4c}}{2\mu},$$

and $T(\mu, c)$ satisfies $T = c^2/(1 - c + c\mu + \mu T)$. Similarly, if $d/n \to c \in (1, \infty)$, then

$$n^{-1}\mathbb{E}\mathrm{tr}((d^{-1}X_n X_n^\top + \mu_n I)^{-1}) \to \tilde{T}(\mu, c) \coloneqq \frac{1 - c - c\mu + \sqrt{(c\mu + c + 1)^2 - 4c}}{2\mu},$$

and $\tilde{T}(\mu, c)$ satisfies $\tilde{T} = c/(c - 1 + c\mu + \mu\tilde{T})$ and $\tilde{T}(\mu, c) = c^2 T(c\mu, c^{-1})$.

Proof. By the Neumann series, $(A + \epsilon I)^{-1} = A^{-1} + \mathcal{O}(\epsilon)$ as $\epsilon \to 0^+$. Therefore,

$$\begin{aligned}
n^{-1}\mathbb{E}\mathrm{tr}\left((d^{-1}X_n^\top X_n + \mu_n I)^{-1}\right) &= n^{-1}\mathbb{E}\mathrm{tr}\left(\left(d^{-1}X_n^\top X_n + \mu I\right)^{-1}\right) + o(1) \\
&= n^{-1}\mathbb{E}\mathrm{tr}\left(\left(\frac{n}{d}n^{-1}X_n^\top X_n + \mu I\right)^{-1}\right) + o(1) \\
&= \frac{d}{n}n^{-1}\mathbb{E}\mathrm{tr}\left(\left(n^{-1}X_n^\top X_n + \frac{d}{n}\mu I\right)^{-1}\right) + o(1).
\end{aligned}$$

By the Marchenko-Pastur Theorem, letting $z = -c\mu$,

$$\begin{aligned}
\frac{d}{n}n^{-1}\mathbb{E}\mathrm{tr}\left((n^{-1}X_n^\top X_n + c\mu I)^{-1}\right) &= \frac{d^2}{n^2}d^{-1}\mathbb{E}\mathrm{tr}\left((n^{-1}X_n^\top X_n + c\mu I)^{-1}\right) \\
&= \frac{d^2}{n^2}d^{-1}\mathbb{E}\mathrm{tr}\left((n^{-1}X_n^\top X_n - zI)^{-1}\right) \\
&\to c^2 \cdot \frac{1 - c - z - \sqrt{(z-c-1)^2 - 4c}}{2cz} \\
&= \frac{c - 1 - c\mu + \sqrt{(c\mu + c + 1)^2 - 4c}}{2\mu}.
\end{aligned}$$

On the other hand, when $c > 1$, letting $\tilde{X}_n = X_n^\top \in \mathbb{R}^{d\times n}$, the Marchenko-Pastur Theorem immediately implies

$$\begin{aligned}
n^{-1}\mathbb{E}\mathrm{tr}\left((d^{-1}X_n X_n^\top + \mu_n I)^{-1}\right) &= n^{-1}\mathbb{E}\mathrm{tr}((d^{-1}\tilde{X}_n^\top \tilde{X}_n + \mu I)^{-1}) + o(1) \\
&\to \frac{c^{-1} - 1 - \mu + \sqrt{(\mu + c^{-1} + 1)^2 - 4c^{-1}}}{2c^{-1}\mu} = \tilde{T}(\mu, c).
\end{aligned}$$

□

Now we turn our attention to the log-determinant, which also depends exclusively on the spectrum. Our method of proof relies on Jacobi's formula, which relates the log-determinant to the trace of the matrix inverse.

Lemma 4 (Log-Determinant)**.** *Let $n, d \to \infty$ such that $d/n \to c \in (0, 1]$ and assume that μ_n is a sequence of real numbers such that $\mu_n \to \mu \in (0, \infty)$ as $n \to \infty$. Then*

$$\frac{1}{n}\mathbb{E}\log\det(d^{-1}X_n^\top X_n + \mu_n I) \to D(\mu, c),$$

where

$$\begin{aligned} D(\mu, c) &\coloneqq (c-1)\log(1-c) - c\log c - c + \int_0^\mu T(t, c)\mathrm{d}t \\ &= \log\left(1 + \frac{T(\mu, c)}{c}\right) - \frac{T(\mu, c)}{c + T(\mu, c)} - c\log\left(\frac{T(\mu, c)}{c}\right). \end{aligned}$$

Similarly, if $d/n \to c \in (1, \infty)$, then

$$\frac{1}{n}\mathbb{E}\log\det(d^{-1}X_n X_n^\top + \mu_n I) \to \tilde{D}(\mu, c),$$

where

$$\begin{aligned} \tilde{D}(\mu, c) &\coloneqq (1-c)\log(c-1) + (c-1)\log c - 1 + \int_0^\mu \tilde{T}(t, c)\mathrm{d}t, \\ &= c\log\left(1 + \frac{\tilde{T}(\mu, c)}{c}\right) - \frac{c\tilde{T}(\mu, c)}{c + \tilde{T}(\mu, c)} - \log\tilde{T}(\mu, c). \end{aligned}$$

Proof. By Jacobi's formula and Taylor's theorem, $\log\det(A + \epsilon I) = \log\det A + \mathcal{O}(\epsilon)$ as $\epsilon \to 0^+$, and so

$$n^{-1}\mathbb{E}\log\det(d^{-1}X_n^\top X_n + \mu_n I) = n^{-1}\mathbb{E}\log\det(d^{-1}X_n^\top X_n + \mu I) + o(1).$$

Furthermore,

$$\begin{aligned} \frac{1}{n}\mathbb{E}\log\det(d^{-1}X_n^\top X_n + \mu I) &= \frac{1}{n}\mathbb{E}\log\det(d^{-1}X_n^\top X_n) + \frac{1}{n}\int_0^\mu \mathbb{E}\mathrm{tr}\left((d^{-1}X_n^\top X_n + tI)^{-1}\right)\mathrm{d}t, \\ &= \frac{1}{n}\mathbb{E}\log\det(d^{-1}X_n^\top X_n) + \int_0^\mu T(t, c)\mathrm{d}t + o(1), \end{aligned}$$

and so it suffices to consider the case $\mu = 0$. Since the log-determinant depends only on the spectrum of X_n, and the spectrum of $n^{-1}X_n^\top X_n$ is asymptotically equivalent to that of $n^{-1}W_n^\top W_n$, where W_n is a Wishart-distributed matrix, it will suffice to consider the limit of $n^{-1}\mathbb{E}\log\det(d^{-1}W_n^\top W_n)$. First, recall that (Bishop & Nasrabadi, 2006, B.81)

$$\begin{aligned} \mathbb{E}\log\det(W_n^\top W_n) &= d\log 2 + \sum_{i=1}^d \psi\left(\frac{n-i+1}{2}\right) \\ &= d\log 2 + n\psi\left(\frac{n}{2}\right) - (n-d)\psi\left(\frac{n-d}{2}\right) \\ &\quad + \mathcal{O}(n^{-1}) + \mathcal{O}(d^{-1}). \end{aligned}$$

Since $\psi(x) = \log x + \mathcal{O}(x^{-1})$, letting $d = [cn]$, there is

$$\begin{aligned} \mathbb{E}\log\det\left(W_n^\top W_n\right) &\sim d\log 2 + n\log\left(\frac{n}{2}\right) - (n-d)\log\left(\frac{n-d}{2}\right) - d \\ &\sim n\log n - (n - cn)\log(n - cn) - cn \\ &\sim n\log n - (1-c)n\log n - (1-c)n\log(1-c) - cn \\ &\sim cn\log n - (1-c)n\log(1-c) - cn. \end{aligned}$$

Therefore,

$$n^{-1}\mathbb{E}\log\det(d^{-1}W_n^\top W_n) \sim \frac{cn\log n - (1-c)n\log(1-c) - cn - cn\log cn}{n}$$
$$\to (c-1)\log(1-c) - c - c\log c,$$

and so

$$\frac{1}{n}\mathbb{E}\log\det(d^{-1}X_n^\top X_n + \mu_n I) \to D(\mu, c) \coloneqq (c-1)\log(1-c) - c\log c - c + \int_0^\mu T(t,c)\mathrm{d}t.$$

To obtain the second equality, we will need to compute the integral term. First, observe that by a change of variables, $\int_0^\mu T(t,c)\mathrm{d}t = \int_0^{c\mu}\tau(t,c)\mathrm{d}t$, where

$$\tau(t,c) = \frac{c-1-t+\sqrt{(t+c+1)^2-4c}}{2t},$$

and $T(t,c) = c\tau(c\mu, c)$. Observe that we can rewrite τ as

$$\tau(t,c) = \frac{(c+1+t)^2 - 4c - (t+1-c)^2}{2t\left[\sqrt{(c+1+t)^2-4c} + (t+1-c)\right]}$$
$$= \frac{2c}{\sqrt{(c+1+t)^2-4c} + (t+1-c)}.$$

Now, let

$$v = v(t) = \frac{c+1+t+\sqrt{(c+1+t)^2-4c}}{2},$$

so that $\tau(t,c) = 2c/(2v-2c) = c/(v-c)$. Note that $v^2 - (c+t+1)v + c = 0$. Differentiating this relation in t, we find

$$2vv' - v - (c+t+1)v' = 0,$$

where $v' = \mathrm{d}v/\mathrm{d}t$, and hence

$$v' = \frac{v}{2v-(c+t+1)}.$$

But since $v^2 + c = (c+t+1)v$,

$$v' = \frac{v}{2v - \frac{v^2+c}{v}} = \frac{v^2}{v^2-c}.$$

Altogether,

$$\int \tau(t,c)\mathrm{d}t = \int \frac{c(v^2-c)}{(v-c)v^2}\mathrm{d}v.$$

From a partial fraction expansion,

$$\frac{c(v^2-c)}{(v-c)v^2} = \frac{A}{v^2} + \frac{B}{v} + \frac{C}{v-c},$$

we find that $c(v^2-c) = A(v-c) + Bv^2 - cBv + Cv^2$, implying that $B + C = c$, $A - cB = 0$ and $-Ac = -c^2$. Therefore, $A = c$, $B = 1$, and $C = c-1$, so

$$\frac{c(v^2-c)}{(v-c)v^2} = \frac{c}{v^2} + \frac{1}{v} + \frac{c-1}{v-c}.$$

Hence, an antiderivative of τ is given by

$$-\frac{c}{v} + \log v + (c-1)\log(v-c).$$

Since $v \to 1$ as $t \to 0$,

$$\int_0^{c\mu}\tau(t,c)\mathrm{d}t = -\frac{c}{v} + \log v + (c-1)\log(v-c) + c - (c-1)\log(1-c).$$

Finally, since $v = c(1+\tau(c\mu,c))/\tau(c\mu,c) = c(c+T(\mu,c))/T(\mu,c)$, the result for $n > d$ follows.

Now we consider the $d > n$ case. Then we have

$$\begin{aligned}
n^{-1}\mathbb{E}\log\det\left(d^{-1}X_nX_n^\top + \mu_n I\right) &= n^{-1}\mathbb{E}\log\det\left(d^{-1}X_nX_n^\top + \mu I\right) + o(1)\\
&= \frac{d}{n}d^{-1}\mathbb{E}\log\det\left(\frac{n}{d}n^{-1}X_nX_n^\top + \mu I\right) + o(1)\\
&= \frac{d}{n}d^{-1}\mathbb{E}\log\det\left(n^{-1}X_nX_n^\top + \frac{d}{n}\mu I\right) + \log\left(\frac{n}{d}\right) + o(1)\\
&= cd^{-1}\mathbb{E}\log\det\left(n^{-1}X_nX_n^\top + c\mu I\right) - \log c + o(1)\\
&\to cD(c\mu, c^{-1}) - \log c.
\end{aligned}$$

From the first expression for $D(\mu, c)$, there is

$$\begin{aligned}
cD(c\mu,c^{-1}) &= c(c^{-1}-1)\log(1-c^{-1}) - \log c^{-1} - 1 + \int_0^{c\mu} cT(t,c^{-1})\mathrm{d}t\\
&= (1-c)\log(c-1) + c\log c - 1 + \int_0^{\mu} c^2T(ct,c^{-1})\mathrm{d}t\\
&= (1-c)\log(c-1) + c\log c - 1 + \int_0^{\mu} \tilde{T}(t,c)\mathrm{d}t.
\end{aligned}$$

Finally, from the second expression for $D(\mu, c)$,

$$\begin{aligned}
cD(c\mu,c^{-1}) &= c\log\left(1+\frac{T(c\mu,c^{-1})}{c^{-1}}\right) - \frac{cT(c\mu,c^{-1})}{c^{-1}+T(c\mu,c^{-1})} - \log\left(\frac{T(c\mu,c^{-1})}{c^{-1}}\right),\\
&= c\log\left(1+\frac{\tilde{T}(\mu,c)}{c}\right) - \frac{c\tilde{T}(\mu,c)}{c+\tilde{T}(\mu,c)} - \log\left(\frac{\tilde{T}(\mu,c)}{c}\right),
\end{aligned}$$

from which the result follows.

□

E KERNELS AND GRAM MATRICES

To extend the results of the previous section to Gram matrices, we rely on the approximation theory developed in El Karoui (2010). For a continuous function $\kappa : \mathbb{R} \to \mathbb{R}$ that is continuously differentiable on $(0, \infty)$, two types of kernels are considered:

(I) **Inner product kernels:** $k(x, y) = \kappa(x^\top y/d)$ for $x, y \in \mathbb{R}^d$, and κ is three-times continuously differentiable in a neighbourhood of zero with $\kappa'(0) > 0$.

(II) **Radial basis kernels:** $k(x, y) = \kappa(\|x - y\|^2/d)$ for $x, y \in \mathbb{R}^d$, and κ is three-times continuously differentiable on $(0, \infty)$ with $\kappa' < 0$.

Let $\|A\|_2$ denote the spectral norm of a matrix A. The following theorem combines Theorems 2.1 and 2.2 in El Karoui (2010).

Theorem 3. *For each $n = 1, 2, \ldots$, let $X_n^1, \ldots, X_n^n$ be independent and identically distributed zero-mean random vectors in $\mathbb{R}^d$ with $\mathrm{Cov}(X_k^i) = \sigma^2 I$ and $\mathbb{E}\|X_k^i\|^{5+\delta} < \infty$ for some $\delta > 0$. For a kernel k of type (I) or (II), consider the Gram matrices $K_X^n \in \mathbb{R}^{n\times n}$ with entries $(K_X^n)_{ij} = k(X_n^i, X_n^j)$. If $n, d \to \infty$ such that $d/n \to c \in (0, \infty)$, then there exists an integer k and a bounded sequence of rank k matrices $C_1, C_2, \ldots$ such that*

$$\|K_X^n - (\alpha d^{-1}XX^\top + \beta I + C_n)\|_2 \to 0,$$

where the constants α, β for cases (I) and (II) are, respectively,

*(I) **Inner product kernels:** $\alpha = \kappa'(0)$, $\beta = \kappa(\sigma^2) - \kappa(0) - \kappa'(0)\sigma^2$;*

*(II) **Radial basis kernels:** $\alpha = -2\kappa'(2\sigma^2)$, $\beta = \kappa(0) + 2\sigma^2\kappa'(2\sigma^2) - \kappa(2\sigma^2)$.*

For the remainder of this section, we assume the hypotheses of Theorem 3, so that $\|K_X^n - (\alpha d^{-1} XX^\top + \beta I)\|_2 \to 0$ for some appropriate $\alpha > 0$ and $\beta \in \mathbb{R}$. To apply Theorem 3 with the results of the previous section, we require the following basic lemma.

Lemma 5. *For any symmetric positive-definite matrices $A, B \in \mathbb{R}^{n\times n}$ and $v > 0$,*

$$\frac{1}{n}|\mathrm{tr}((A+vI)^{-1}) - \mathrm{tr}((B+vI)^{-1})| \le \frac{\|A-B\|_2}{v^2}$$
$$\frac{1}{n}|\log\det(A+vI) - \log\det(B+vI)| \le \frac{\|A-B\|_2}{v}.$$

Proof. Let $\lambda_1(A) \ge \cdots \ge \lambda_n(A)$ and $\lambda_1(B) \ge \cdots \ge \lambda_n(B)$ denote the eigenvalues of A and B, respectively, in decreasing order. Recall from Weyl's perturbation theorem (see Corollary III.2.6 of Bhatia (2013)) that $\max_{i=1,\ldots,n} |\lambda_i(A) - \lambda_i(B)| \le \|A-B\|_2$. By the Mean Value Theorem, for any $x, y > 0$, $|(x+v)^{-1} - (y+v)^{-1}| \le v^{-2}|x-y|$. Therefore,

$$\begin{aligned}\frac{1}{n}|\mathrm{tr}((A+vI)^{-1}) - \mathrm{tr}((B+vI)^{-1})| &= \frac{1}{n}\left|\sum_{i=1}^n \frac{1}{\lambda_i(A)+v} - \frac{1}{\lambda_i(B)+v}\right| \\ &\le \frac{1}{v^2}\max_{i=1,\ldots,n}|\lambda_i(A) - \lambda_i(B)| \\ &\le \frac{1}{v^2}\|A-B\|_2.\end{aligned}$$

Similarly, the Mean Value Theorem implies that for any $x, y > 0$, $|\log(x+v) - \log(y+v)| \le v^{-1}|x-y|$, and so

$$\begin{aligned}\frac{1}{n}|\log\det(A+vI) - \log\det(B+vI)| &= \frac{1}{n}\left|\sum_{i=1}^n \log(\lambda_i(A)+v) - \log(\lambda_i(B)+v)\right| \\ &\le \frac{1}{v}\max_{i=1,\ldots,n}|\lambda_i(A) - \lambda_i(B)| \\ &\le \frac{1}{v}\|A-B\|_2.\end{aligned}$$

□

Combining Theorem 3 and Lemma 5 with Lemmas 3 and 4 yields the following corollary.

Corollary 1. *Under the assumptions of Theorem 3, if μ_n is a sequence of positive real numbers such that $\mu_n \to \mu \in (0,\infty)$ as $n \to \infty$, then*

$$\frac{1}{n}\mathbb{E}\mathrm{tr}((K_X^n + \mu_n I)^{-1}) \to \begin{cases} \frac{1-c}{\beta+\mu} + \frac{1}{\alpha}T\left(\frac{\beta+\mu}{\alpha}, c\right) & \text{if } c < 1 \\ \frac{1}{\alpha}\tilde{T}\left(\frac{\beta+\mu}{\alpha}, c\right) & \text{if } c > 1, \end{cases}$$
$$\frac{1}{n}\mathbb{E}\log\det(K_X^n + \mu_n I) \to \begin{cases} D\left(\frac{\beta+\mu}{\alpha}, c\right) + (1-c)\log\left(\frac{\beta+\mu}{\alpha}\right) + \log\alpha & \text{if } c < 1 \\ \tilde{D}\left(\frac{\beta+\mu}{\alpha}, c\right) + \log\alpha & \text{if } c > 1. \end{cases}$$

Proof. First consider the $c > 1$ case. Combining Theorem 3 and Lemma 5, and noting that finite rank perturbations do not affect the limiting spectrum (El Karoui, 2010, Lemma 2.1), we find that

$$\begin{aligned}\frac{1}{n}\mathbb{E}\mathrm{tr}\left(K_X^n + \mu_n I\right)^{-1} &= \frac{1}{n}\mathbb{E}\mathrm{tr}\left(\alpha d^{-1}XX^\top + \beta I + \mu_n I\right)^{-1} + o(1) \\ &= \frac{1}{\alpha n}\mathbb{E}\mathrm{tr}\left(d^{-1}XX^\top + \frac{\beta+\mu}{\alpha}I\right)^{-1} + o(1) \\ &\to \frac{1}{\alpha}\tilde{T}\left(\frac{\beta+\mu}{\alpha}, c\right).\end{aligned}$$

Similarly, since $XX^\top \in \mathbb{R}^{n\times n}$,

$$\begin{aligned}\frac{1}{n}\mathbb{E}\log\det\left(K_X^n+\mu_n I\right) &= \frac{1}{n}\mathbb{E}\log\det\left(\alpha d^{-1}XX^\top+\beta I+\mu I\right)+o(1)\\ &= \frac{1}{n}\mathbb{E}\log\det\left(d^{-1}XX^\top+\frac{\beta+\mu}{\alpha}I\right)+\log\alpha+o(1)\\ &\to \tilde{D}\left(\frac{\beta+\mu}{\alpha},c\right)+\log\alpha.\end{aligned}$$

For the $c<1$ case, from the Woodbury matrix identity (Pozrikidis, 2014, B.1.2),

$$\begin{aligned}\text{tr}\left((\eta_1 XX^\top+\eta_2 I)^{-1}\right) &= \frac{n}{\eta_2}-\text{tr}\left(\frac{\eta_1}{\eta_2}X\left(\eta_2 I+\eta_1 X^\top X\right)^{-1}X^\top\right)\\ &= \frac{n}{\eta_2}-\frac{1}{\eta_2}\text{tr}\left(\left(\eta_2 I+\eta_1 X^\top X\right)^{-1}\eta_1 X^\top X\right)\\ &= \frac{n}{\eta_2}-\frac{1}{\eta_2}\text{tr}\left(I-\eta_2\left(\eta_2 I+\eta_1 X^\top X\right)^{-1}\right)\\ &= \frac{n-d}{\eta_2}+\text{tr}\left((\eta_1 X^\top X+\eta_2 I)^{-1}\right).\end{aligned}$$

Therefore,

$$\begin{aligned}\frac{1}{n}\mathbb{E}\text{tr}\left(K_X^n+\mu_n I\right)^{-1} &= \frac{1}{\alpha n}\mathbb{E}\text{tr}\left(d^{-1}XX^\top+\frac{\beta+\mu}{\alpha}I\right)^{-1}+o(1)\\ &= \frac{1-\frac{d}{n}}{\beta+\mu}+\frac{1}{\alpha n}\mathbb{E}\text{tr}\left(d^{-1}X^\top X+\frac{\beta+\mu}{\alpha}I\right)^{-1}+o(1)\\ &\to \frac{1-c}{\beta+\mu}+\frac{1}{\alpha}T\left(\frac{\beta+\mu}{\alpha},c\right).\end{aligned}$$

Finally, from Sylvester's determinant theorem (Pozrikidis, 2014, B.1.15),

$$\begin{aligned}\log\det(\eta_1 XX^\top+\eta_2 I) &= \log\det\left(\frac{\eta_1}{\eta_2}XX^\top+I\right)+n\log\eta_2\\ &= \log\det\left(\frac{\eta_1}{\eta_2}X^\top X+I\right)+n\log\eta_2\\ &= \log\det\left(\eta_1 X^\top X+\eta_2 I\right)+(n-d)\log\eta_2.\end{aligned}$$

Therefore,

$$\begin{aligned}\frac{1}{n}\mathbb{E}\log\det\left(K_X^n+\mu_n I\right) &= \frac{1}{n}\mathbb{E}\log\det\left(d^{-1}XX^\top+\frac{\beta+\mu}{\alpha}I\right)+\log\alpha+o(1)\\ &= \frac{1}{n}\mathbb{E}\log\det\left(d^{-1}X^\top X+\frac{\beta+\mu}{\alpha}I\right)\\ &\qquad+\left(1-\frac{d}{n}\right)\log\left(\frac{\beta+\mu}{\alpha}\right)+\log\alpha+o(1)\\ &\to D\left(\frac{\beta+\mu}{\alpha},c\right)+(1-c)\log\left(\frac{\beta+\mu}{\alpha}\right)+\log\alpha.\end{aligned}$$

□

F Proofs of Main Results

With the underlying random matrix theory in place, we can begin to prove our main result in Theorem 1. Throughout this section, we assume the conditions of Theorem 1, that is, we let $X_1, X_2, \ldots$ be independent and identically distributed zero-mean random vectors in $\mathbb{R}^d$ with unit covariance, satisfying $\mathbb{E}\|X_i\|^{5+\delta}<+\infty$ for some $\delta>0$. For each $n=1,2,\ldots$, let

$$\mathcal{F}_n^\gamma = \tfrac{1}{2}\lambda Y^\top(K_X+\lambda\gamma I)^{-1}Y+\tfrac{1}{2}\log\det(K_X+\lambda\gamma I)-\tfrac{n}{2}\log\left(\tfrac{\lambda}{2\pi}\right).$$

where $K_X\in\mathbb{R}^{n\times n}$ satisfies $K_X^{ij}=k(X_i,X_j)$ and $Y=(Y_i)_{i=1}^n$, with each $Y_i\sim\mathcal{N}(0,1)$.

Proposition 2 (El Karoui-Marchenko-Pastur Limit of the Bayes Free Energy)**.** *Assuming that $n, d \to \infty$ such that $d/n \to c \in (0, \infty)$, there is $n^{-1}\mathbb{E}\mathcal{F}_n^\gamma \to \mathcal{F}_\infty^\gamma$ where for $c < 1$,*

$$\mathcal{F}_\infty^\gamma = \frac{\lambda}{2}\left(\frac{1-c}{\beta+\gamma\lambda} + \frac{1}{\alpha}T\left(\frac{\beta+\gamma\lambda}{\alpha}, c\right)\right) - \frac{1}{2}\log\left(\frac{\lambda}{2\pi\alpha}\right) + \frac{1}{2}D\left(\frac{\beta+\gamma\lambda}{\alpha}, c\right) + \frac{1}{2}(1-c)\log\left(\frac{\beta+\gamma\lambda}{\alpha}\right),$$

and for $c > 1$,

$$\mathcal{F}_\infty^\gamma = \frac{\lambda}{2\alpha}\tilde{T}\left(\frac{\beta+\gamma\lambda}{\alpha}, c\right) - \frac{1}{2}\log\left(\frac{\lambda}{2\pi\alpha}\right) + \frac{1}{2}\tilde{D}\left(\frac{\beta+\gamma\lambda}{\alpha}, c\right).$$

Proof. Recalling that $\mathbb{E}[Y^\top AY] = \mathrm{tr}(A)$ for any $A \in \mathbb{R}^{n\times n}$, since K_X is independent of Y,

$$\frac{1}{n}\mathbb{E}\mathcal{F}_n^\gamma = \frac{\lambda}{2n}\mathbb{E}\mathrm{tr}((K_X + \lambda\gamma I)^{-1}) + \frac{1}{2n}\mathbb{E}\log\det(K_X + \lambda\gamma I) - \frac{1}{2}\log\left(\frac{\lambda}{2\pi}\right).$$

The result follows by a direct application of Corollary 1. □

Proposition 3 (Optimal Temperature in the Bayes Free Energy)**.** *Assume that $\lambda = \mu/\gamma$ for some fixed $\mu > 0$. The limiting Bayes free energy $\mathcal{F}_\infty^\gamma$ is minimized in γ at*

$$\gamma^* = \frac{\mu}{2(\beta+\mu)}[1 - c - c(\tfrac{\beta+\mu}{\alpha}) + \sqrt{(c(\tfrac{\beta+\mu}{\alpha}) + c + 1)^2 - 4c}].$$

Proof. First consider the case $c < 1$. If $\lambda = \mu/\gamma$, then

$$\mathcal{F}_\infty^\gamma = \frac{\mu}{2\gamma}\left(\frac{1-c}{\beta+\mu} + \frac{1}{\alpha}T\left(\frac{\beta+\mu}{\alpha}, c\right)\right) - \frac{1}{2}\log\left(\frac{\mu}{2\pi\gamma\alpha}\right) \frac{1}{2}D\left(\frac{\beta+\mu}{\alpha}, c\right) + \frac{1}{2}(1-c)\log\left(\frac{\beta+\mu}{\alpha}\right).$$

Note that as $\gamma \to 0^+$ or $\gamma \to \infty$, $\mathcal{F}_\infty^\gamma$, so if there exists only one point γ^* where that $\partial\mathcal{F}_\infty^\gamma/\partial\gamma = 0$, then by Fermat's Theorem, γ^* is the unique global minimizer of $\mathcal{F}_\infty^\gamma$. For μ fixed, we may differentiate in γ to find that

$$\frac{\partial\mathcal{F}_\infty^\gamma}{\partial\gamma} = -\frac{\mu}{2\gamma^2}\left(\frac{1-c}{\beta+\mu} + \frac{1}{\alpha}T\left(\frac{\beta+\mu}{\alpha}, c\right)\right) + \frac{1}{2\gamma}.$$

Solving $\partial\mathcal{F}_\infty^\gamma/\partial\gamma = 0$ for γ, the optimal

$$\gamma^* = \mu\left(\frac{1-c}{\beta+\mu} + \frac{1}{\alpha}T\left(\frac{\beta+\mu}{\alpha}, c\right)\right).$$

Simplifying,

$$\frac{1-c}{\beta+\mu} + \frac{1}{\alpha}\frac{c - 1 - c(\frac{\beta+\mu}{\alpha}) + \sqrt{(c(\frac{\beta+\mu}{\alpha}) + c + 1)^2 - 4c}}{2(\frac{\beta+\mu}{\alpha})} = \frac{1 - c - c(\frac{\beta+\mu}{\alpha}) + \sqrt{(c(\frac{\beta+\mu}{\alpha}) + c + 1)^2 - 4c}}{2(\beta+\mu)},$$

which implies the result for $c < 1$. On the other hand, for $c > 1$,

$$\mathcal{F}_\infty^\gamma = \frac{\mu}{2\gamma\alpha}\tilde{T}\left(\frac{\beta+\mu}{\alpha}, c\right) - \frac{1}{2}\log\left(\frac{\mu}{2\pi\alpha\gamma}\right) + \frac{1}{2}\tilde{D}\left(\frac{\beta+\mu}{\alpha}, c\right),$$

and once again, as $\gamma \to 0^+$ or $\gamma \to \infty$, $\mathcal{F}_\infty^\gamma \to \infty$, so a unique critical point is the unique global minimizer of $\mathcal{F}_\infty^\gamma$. For μ fixed, we differentiate in γ to find

$$\frac{\partial\mathcal{F}_\infty^\gamma}{\partial\gamma} = -\frac{\mu}{2\gamma^2\alpha}\tilde{T}\left(\frac{\beta+\mu}{\alpha}, c\right) + \frac{1}{2\gamma}.$$

Solving $\partial \mathcal{F}_\infty^\gamma / \partial \gamma = 0$ for γ, the optimal

$$\gamma^* = \frac{\mu}{\alpha} \tilde{T}\left(\frac{\beta+\mu}{\alpha}, c\right).$$

Simplifying,

$$\frac{1}{\alpha}\tilde{T}\left(\frac{\beta+\mu}{\alpha}, c\right) = \frac{1 - c - c(\frac{\beta+\mu}{\alpha}) + \sqrt{(c(\frac{\beta+\mu}{\alpha}) + c + 1)^2 - 4c}}{2(\beta+\mu)},$$

which implies the result for $c > 1$. □

In the sequel, we assume that the kernel itself depends on λ in such a way that $\beta = \beta_0 \lambda$ for some $0 < \beta_0 < 1$. Let $\gamma_0 = \gamma + \beta_0$ and $\mu = \lambda\gamma_0/\alpha$. For $c < 1$, the limiting Bayes free energy satisfies

$$\begin{aligned}
\mathcal{F}_\infty^\gamma &= \frac{1}{2\gamma_0}(1 - c + \mu T(\mu, c)) - \frac{1}{2}\log\left(\frac{\mu}{2\pi\gamma_0}\right) + \frac{1}{2}D(\mu, c) + \frac{1}{2}(1-c)\log\mu, \\
&= \frac{1}{2\gamma_0}(1 - c + \mu T(\mu, c)) - \frac{1}{2}\log\left(\frac{1}{2\pi\gamma_0}\right) + \frac{1}{2}D(\mu, c) - \frac{c}{2}\log\mu.
\end{aligned}$$

and for $c > 1$,

$$\mathcal{F}_\infty^\gamma = \frac{\mu}{2\gamma_0}\tilde{T}(\mu, c) - \frac{1}{2}\log\left(\frac{\mu}{2\pi\gamma_0}\right) + \frac{1}{2}\tilde{D}(\mu, c).$$

Proposition 4 (Optimal Regularization in the Bayes Free Energy)**.** *The limiting Bayes free energy $\mathcal{F}_\infty^\gamma$ is minimized in λ at*

$$\lambda^* = \frac{\alpha[(c+1)\gamma_0 + \sqrt{(c-1)^2 + 4c\gamma_0^2}]}{c(1-\gamma_0^2)}.$$

Proof. Since $\mathcal{F}_\infty^\gamma$ is smooth for $\lambda \in (0, \infty)$ (and therefore $\mu \in (0, \infty)$), Fermat's theorem implies that any optimal temperature λ^* must be a critical point of $\mathcal{F}_\infty^\gamma$ in $(0, \infty)$. First, consider the case where $c < 1$. Differentiating $\mathcal{F}_\infty^\gamma$ with respect to μ,

$$\frac{\partial \mathcal{F}_\infty^\gamma}{\partial \mu} = \frac{1}{2\gamma_0}\frac{\partial}{\partial \mu}(\mu T(\mu, c)) - \frac{c}{2\mu} + \frac{1}{2}T(\mu, c).$$

Letting $U(\mu, c) = \mu T(\mu, c)$ and $U' = \frac{\partial U}{\partial \mu}$,

$$\frac{\partial \mathcal{F}_\infty^\gamma}{\partial \mu} = \frac{1}{2\mu}\left(\frac{\mu}{\gamma_0}U' + U - c\right). \tag{8}$$

Noting that

$$U(\mu, c) = \frac{c - 1 - c\mu + \sqrt{(c\mu + c + 1)^2 - 4c}}{2},$$

and

$$U' = -\frac{c}{2} + \frac{c(c\mu + c + 1)}{2\sqrt{(c\mu + c + 1)^2 - 4c}} = c \cdot \frac{c\mu + c + 1 - \sqrt{(c\mu + c + 1)^2 - 4c}}{2\sqrt{(c\mu + c + 1)^2 - 4c}},$$

and so $U'\sqrt{(c\mu + c + 1)^2 - 4c} = c \cdot (c - U)$. Therefore, substituting into (8) reveals

$$\frac{\partial \mathcal{F}_\infty^\gamma}{\partial \mu} = \frac{1}{2c\mu}\left(\frac{c\mu}{\gamma_0} - \sqrt{(c\mu + c + 1)^2 - 4c}\right)U'.$$

Since $U' > 0$, $\partial \mathcal{F}_\infty^\gamma / \partial \mu = 0$ if and only if

$$\frac{c\mu}{\gamma_0} = \sqrt{(c\mu + c + 1)^2 - 4c}. \tag{9}$$

This occurs when

$$c^2(1 - \gamma_0^2)\mu^2 - 2c\mu(c+1)\gamma_0^2 - (c-1)^2\gamma_0^2 = 0. \tag{10}$$

If $\gamma_0 \geq 1$, then no positive solutions exist for μ. On the other hand, if $\gamma < 1$, then only one positive solution exists, and is given by

$$\begin{aligned}\mu^* &= \frac{2c(c+1)\gamma_0^2 + \sqrt{4c^2(c+1)^2\gamma_0^4 + 4c^2(1-\gamma_0^2)(c-1)^2\gamma_0^2}}{2c^2(1-\gamma_0^2)} \\ &= \frac{2c(c+1)\gamma_0^2 + 2c\gamma_0\sqrt{[(c+1)^2-(c-1)^2]\gamma_0^2 + (c-1)^2}}{2c^2(1-\gamma_0^2)} \\ &= \frac{(c+1)\gamma_0^2 + \gamma_0\sqrt{(c-1)^2 + 4c\gamma_0^2}}{c(1-\gamma_0^2)}.\end{aligned}$$

Next, consider the case $c > 1$. Differentiating $\mathcal{F}_\infty^\gamma$ with respect to μ, we seek

$$\frac{\partial \mathcal{F}_\infty^\gamma}{\partial \mu} = \frac{1}{2\gamma}\frac{\partial}{\partial \mu}(\mu\tilde{T}(\mu, c)) + \frac{1}{2}\tilde{T}(\mu, c) - \frac{1}{2\mu} = 0,$$

or, equivalently,

$$\frac{\mu}{\gamma}\frac{\partial}{\partial \mu}(\mu\tilde{T}(\mu, c)) + \mu\tilde{T}(\mu, c) - 1 = 0. \tag{11}$$

Letting $\tilde{U} = \mu\tilde{T}$ and $\tilde{U}' = \frac{\partial \tilde{U}}{\partial \mu}$, we require $\frac{\mu}{\gamma}U' + U - 1 = 0$. But since

$$\tilde{U} = \frac{1 - c - c\mu + \sqrt{(c\mu + c + 1)^2 - 4c}}{2},$$

and

$$\tilde{U}' = \frac{\partial \tilde{U}}{\partial \mu} = \frac{c(c + c\mu + 1 - \sqrt{(c\mu + c + 1)^2 - 4c})}{2\sqrt{(c\mu + c + 1)^2 - 4c}},$$

we find that $\tilde{U}'\sqrt{(c\mu + c + 1)^2 - 4c} = c(1 - \tilde{U})$. Substituting this relation into (9), we obtain

$$\frac{\partial \mathcal{F}_\infty^\gamma}{\partial \mu} = \frac{1}{2\mu c}\left(\frac{c\mu}{\gamma} - \sqrt{(c\mu + c + 1)^2 - 4c}\right)\tilde{U}' = 0,$$

and since $U' > 0$, an optimal μ^* occurs if and only if (9) holds. The rest of the proof proceeds as in the $c < 1$ case. □

Proposition 5 (Monotonicity in the Bayes Free Energy)**.** *The limiting Bayes free energy $\mathcal{F}_\infty^\gamma$ at $\lambda = \lambda^*$ decreases monotonically in $c \in (0, \infty)$.*

Proof. First, we treat the $c < 1$ case. Using the closed form expression for $D(\mu, c)$ in Lemma 4,

$$\begin{aligned}\mathcal{F}_\infty^\gamma = \frac{1}{2\gamma_0}(1 - c + \mu T) + \frac{1}{2}\log(2\pi\gamma_0) - \frac{c}{2}\log\mu \\ + \frac{1}{2}\left[\log\left(1 + \frac{T}{c}\right) - \frac{T}{c+T} - c\log\left(\frac{T}{c}\right)\right].\end{aligned}$$

Note that, at the optimal μ^*, $\frac{\mathrm{d}}{\mathrm{d}c}\mathcal{F}_\infty^\gamma = \frac{\partial}{\partial c}\mathcal{F}_\infty^\gamma + \frac{\partial}{\partial \mu}\mathcal{F}_\infty^\gamma \cdot \frac{\partial \mu^*}{\partial c} = \frac{\partial}{\partial c}\mathcal{F}_\infty^\gamma$. Therefore,

$$2\frac{\mathrm{d}\mathcal{F}_\infty^\gamma}{\mathrm{d}c} = \left(\frac{T}{(T+c)^2} + \frac{\mu}{\gamma_0} - \frac{c}{T}\right)\frac{\partial T}{\partial c} + 1 - \frac{1}{\gamma_0} - \frac{T^2}{c(T+c)^2} + \log\left(\frac{c}{\mu T}\right).$$

Differentiating T in c, we find that

$$\begin{aligned}\frac{\partial T}{\partial c} &= \frac{1-\mu}{2\mu} + \frac{1}{2\mu}\left(\frac{(c\mu + c + 1)(\mu + 1) - 2}{\sqrt{(c\mu + c + 1)^2 - 4c}}\right) \\ &= \frac{(c\mu + c + 1)(\mu + 1) - (\mu - 1)\sqrt{(c\mu + c + 1)^2 - 4c} - 2}{2\mu\sqrt{(c\mu + c + 1)^2 - 4c}} \\ &= \frac{2c - (\mu - 1)T}{\sqrt{(c\mu + c + 1)^2 - 4c}}.\end{aligned}$$

Since $c\mu^*/\gamma_0 = \sqrt{(c\mu^*+c+1)^2-4c}$, at the optimal μ^*,

$$\frac{\partial T}{\partial c} = \gamma_0 \cdot \frac{2c-\mu T+T}{c\mu}.$$

Note that for any $c>0$,

$$\mu T = \frac{c-1-c\mu+\sqrt{(c\mu+c+1)^2-4c}}{2} < \frac{c-1-c\mu+c\mu+c+1}{2} = c.$$

Recalling that $\log x < x-1$ for any $x>1$, $\log(c/(\mu T)) < c/(\mu T)-1$. Therefore, $2\frac{\mathrm{d}\mathcal{F}_\infty^\gamma}{\mathrm{d}c} < M$, where

$$M = \left(\frac{T}{(T+c)^2}+\frac{\mu}{\gamma_0}-\frac{c}{T}\right)\frac{\gamma_0(2c-\mu T+T)}{c\mu} - \frac{1}{\gamma_0} - \frac{T^2}{c(T+c)^2} + \frac{c}{\mu T}.$$

Since $T = (c-1-c\mu^*+c\mu^*/\gamma_0)/(2\mu^*)$ at the optimal μ^*, after several calculations, we find that

$$M = Q(\mu^*,c,\gamma_0)\frac{c^2(1-\gamma_0^2)(\mu^*)^2 - 2c\mu^*(c+1)\gamma_0^2 - (c-1)^2\gamma_0^2}{2c\gamma_0\mu^*(c\gamma_0\mu^*-c\gamma_0+c\mu^*+\gamma_0)(c\gamma_0\mu^*+c\gamma_0-c\mu^*-\gamma_0)},$$

where

$$\begin{aligned}Q(\mu,c,\gamma) = (2(c-1)\gamma+c\mu)(c\mu^2+2c\mu+c+\mu-1)\gamma^2 \\ + c\mu(c\mu+c+\mu-1)\gamma^2 - (\mu+1)(c\mu+(c-1)\gamma)^2.\end{aligned}$$

In particular, by (10), at $\mu=\mu^*$, $M=0$, and hence, $\frac{\mathrm{d}}{\mathrm{d}c}\mathcal{F}_\infty^\gamma < 0$.

Next, we treat the $c>1$ case. Using the closed form expression in Lemma 4,

$$\mathcal{F}_\infty^\gamma = \frac{\mu}{2\gamma_0}\tilde{T} - \frac{1}{2}\log\left(\frac{\mu}{2\pi\gamma_0}\right) + \frac{1}{2}c\log(c+\tilde{T}) - \frac{1}{2}c\log c - \frac{1}{2}\frac{c\tilde{T}}{c+\tilde{T}} - \frac{1}{2}\log\tilde{T}.$$

Differentiating in c at the optimal μ^*,

$$\begin{aligned}2\frac{\mathrm{d}\mathcal{F}_\infty^\gamma}{\mathrm{d}c} = 2\frac{\partial\mathcal{F}_\infty^\gamma}{\partial c} = \left(\frac{\mu}{\gamma_0}-\frac{1}{\tilde{T}}+\frac{c\tilde{T}}{(c+\tilde{T})^2}\right)\frac{\partial\tilde{T}}{\partial c} \\ + \log(c+\tilde{T}) - \log c - 1 + \frac{c\tilde{T}}{(c+\tilde{T})^2} + \frac{c^2}{(c+\tilde{T})^2} - \frac{\tilde{T}^2}{(c+\tilde{T})^2}.\end{aligned}$$

Differentiating $\tilde{T}$ in c, we find that

$$\begin{aligned}\frac{\partial\tilde{T}}{\partial c} &= \frac{-1-\mu+\frac{(c\mu+c+1)(\mu+1)-2}{\sqrt{(c\mu+c+1)^2-4c}}}{2\mu} \\ &= \frac{-(\mu+1)\sqrt{(c\mu+c+1)^2-4c}+(c\mu+c+1)(\mu+1)-2}{2\mu\sqrt{(c\mu+c+1)^2-4c}} \\ &= \frac{1}{\sqrt{(c\mu+c+1)^2-4c}}\cdot\left[(\mu+1)\frac{c\mu+c+1-\sqrt{(c\mu+c+1)^2-4c}}{2\mu}-\frac{1}{\mu}\right] \\ &= \frac{1}{\sqrt{(c\mu+c+1)^2-4c}}\cdot\left[(\mu+1)\left(\frac{1}{\mu}-\tilde{T}\right)-\frac{1}{\mu}\right] \\ &= \frac{1-\mu\tilde{T}-\tilde{T}}{\sqrt{(c\mu+c+1)^2-4c}}.\end{aligned}$$

Since $c\mu^*/\gamma_0 = \sqrt{(c\mu^*+c+1)^2-4c}$, it follows that

$$\frac{\partial\tilde{T}}{\partial c} = \gamma\cdot\frac{1-\mu T-T}{c\mu}.$$

Note that for any $c > 0$, $\tilde{T} < c$, and so $\log(1 + \tilde{T}/c) < \tilde{T}/c$. Therefore, $2\frac{\mathrm{d}\mathcal{F}_\infty^\gamma}{\mathrm{d}c} < M$ where

$$M = \frac{\gamma_0}{c\mu}\left(\frac{\mu}{\gamma_0} - \frac{1}{\tilde{T}} + \frac{c\tilde{T}}{(c+\tilde{T})^2}\right) + \frac{\tilde{T}}{c} - 1 + \frac{c\tilde{T}}{(c+\tilde{T})^2} + \frac{c^2}{(c+\tilde{T})^2} - \frac{\tilde{T}^2}{(c+\tilde{T})^2}.$$

Since $\tilde{T} = (1 - c - c\mu^* + c\mu^*/\gamma_0)/(2\mu^*)$ at the optimal μ^*, after several calculations, we find that

$$M = -Q(\mu^*, c, \gamma_0)\frac{c^2(1-\gamma_0^2)(\mu^*)^2 - 2c\mu^*(c+1)\gamma_0^2 - (c-1)^2\gamma_0^2}{2c\gamma_0\mu^*(c\gamma_0\mu^* - c\gamma_0 + c\mu^* + \gamma_0)(c\gamma_0\mu^* + c\gamma_0 - c\mu^* - \gamma_0)},$$

where

$$\begin{aligned} Q(\mu, c, \gamma) = \mu(c\mu+\gamma)^2 + 2c(\mu+1)^2(c-1)\gamma^3 + 2(c-1)(\mu-1)\gamma^3 \\ - 2c^2\gamma^2\mu(\mu+1) - \mu c^2\gamma^2(\mu+1)^2 - 2c\gamma^2\mu(\mu-1). \end{aligned}$$

In particular, since the numerator for M is always zero, it follows that $\frac{\mathrm{d}\mathcal{F}_\infty^\gamma}{\mathrm{d}c} < 0$. □

Theorem 1 follows immediately from Propositions 2, 3, 4, and 5.

Proof of Proposition 1. Under the stated hypotheses, let $\delta(\lambda, \gamma) = c_2(\lambda, \gamma)/c_1(\gamma)$ and $\bar{E}(c) = E(c) + c_3(\gamma)/c_1(\gamma)$. Then

$$\begin{aligned} |\mathcal{L}/c_1 - \bar{E}| &\le |\mathbb{E}\|\bar{f}(\boldsymbol{x}) - \boldsymbol{y}\|^2 - E| + \delta(\lambda(\gamma), \gamma)\mathbb{E}\mathrm{tr}(\Sigma(\boldsymbol{x})) \\ &\le |\mathbb{E}\|\bar{f}(\boldsymbol{x}) - \boldsymbol{y}\|^2 - E| + \delta(\lambda(\gamma), \gamma)m\mathbb{E}k(x, x). \end{aligned}$$

For an arbitrary $\epsilon > 0$, let N be sufficiently large so that for any $n > N$ and $d = d(n)$, $|\mathbb{E}\|\bar{f}(\boldsymbol{x}) - \boldsymbol{y}\|^2 - E| \le \epsilon/2$. Similarly, let γ_0 be sufficiently small so that for any $0 < \gamma < \gamma_0$, $\delta(\lambda(\gamma), \gamma) < \epsilon/(2m\mathbb{E}k(x, x))$. Then $|\mathcal{L}/c_1 - \bar{E}| < \epsilon$, and the result follows. □

G Details of Experiments

In each figure shown throughout this work, a performance metric has been calculated for varying dataset size n, input dimension d, and hyperparameters γ, λ. For experiments involving synthetic data, $X \in \mathbb{R}^{n\times d}$ has iid rows drawn from $\mathcal{N}(0, \Sigma)$, and $Y = (Y_i)_{i=1}^n$ is comprised of iid samples from $\mathcal{N}(0, \sigma^2)$ (where $\Sigma = I$ and $\sigma = 1$ unless specified otherwise). For PPL2 and PPNLL, the expectation is computed over iid scalar test points $x, y \sim \mathcal{N}(0, 1)$. Runs are averaged over a number of iterations, and 95% confidence intervals (under the central limit theorem) are highlighted. In Table 2 we present the parameters used for each figure.

Figure	n	d	γ	λ	Kernel	Iterations	Notes
2UL	300	$[100, 1000]$	$[0.01, 0.99]$	λ^*	linear	5	
2UR	300	$[100, 1000]$	$[0.01, 0.99]$	0.01	linear	5	
2BL	300	$[100, 1000]$	$[0.01, 0.99]$	λ^*	Gaussian	5	
2BR	300	$[100, 1000]$	$[0.01, 0.99]$	0.01	Gaussian	5	
3L	900	$[180, 3003]$	$[0.01, 0.99]$	λ^*	linear	100	`CIFAR10` dataset
3R	900	$[180, 3003]$	$[0.01, 0.99]$	λ^*	Gaussian	100	`CIFAR10` dataset
4	50	$[10, 368]$	$[0.01, 0.99]$	λ^*	linear	100	`CT Slices` dataset (augmented)
5L	100	$[10, 1000]$	$[0.001, 0.99]$	$0.01/\gamma$	linear	100	
5R	100	$[10, 1000]$	$[0.001, 0.99]$	λ^*	linear	100	
6L	900	$[180, 3003]$	$[0.01, 0.99]$	λ^*	linear	100	`CIFAR10` dataset
6R	900	$[180, 3003]$	$[0.01, 0.99]$	λ^*	linear	100	`CIFAR10` dataset
7L	50	$[10, 368]$	$[0.01, 0.99]$	λ^*	linear	100	`CT Slices` dataset
7R	50	$[10, 368]$	$[0.01, 0.99]$	λ^*	Gaussian	100	`CT Slices` dataset
8L	50	$[10, 368]$	$[0.01, 0.99]$	$0.01/\gamma$	linear	100	`CT Slices` dataset
8R	50	$[10, 368]$	$[0.01, 0.99]$	λ^*	linear	100	`CT Slices` dataset
9L	175	$[35, 706]$	$[0.01, 0.99]$	λ^*	linear	100	`MNIST` dataset
9R	175	$[35, 706]$	$[0.01, 0.99]$	λ^*	Gaussian	100	`MNIST` dataset
10L	175	$[35, 706]$	$[0.01, 0.99]$	$0.01/\gamma$	linear	100	`MNIST` dataset
10R	175	$[35, 706]$	$[0.01, 0.99]$	λ^*	linear	100	`MNIST` dataset
11	100	$[30, 300]$	$[0.001, 0.1]$	λ_{opt}	linear	10000	1000 iterations for $d > n$
12	100	$[10, 1000]$	γ_{opt}	$[0.1/\gamma_{opt}, 10/\gamma_{opt}]$	linear	100	
13L	300	$[100, 1000]$	$[0.01, 0.99]$	λ^*	linear	5	$\|\theta_0\| = 1$
13C	300	$[100, 1000]$	$[0.01, 0.99]$	λ^*	linear	5	$\|\theta_0\| = \sqrt{d}$
13R	300	$[100, 1000]$	$[0.01, 0.99]$	λ^*	linear	5	$\|\theta_0\| = \sqrt{n}$
14	300	$[100, 1000]$	0.01	λ^*	Matérn	5	$\nu \in [0.5, 100]$
15	300	$[100, 1000]$	$[0.01, 0.99]$	λ^*	linear	5	$\Sigma = \text{diag}((10)_{i=1}^{d/2}, (1/10)_{i=1}^{d/2})$
16L	[300,3000]	$2^{10(1-\xi)}n^{\xi}$	0.01	λ^*	linear	5	$\xi \in [0.1, 1.9]$
16R	[300,3000]	$2^{10(1-\xi)}n^{\xi}$	0.01	λ^*	Gaussian	5	$\xi \in [0.1, 1.9]$
17L	300	$[100, 1000]$	0.1	λ^*	linear	5	$\sigma^2 \in [0.1, 10]$
17R	300	$[100, 1000]$	0.01	λ^*	linear	5	$\sigma^2 \in [0.1, 10]$
18	100	$[10, 1000]$	$[0.001, 0.99]$	$0.01/\gamma$	Gaussian	100	
19	100	$[10, 1000]$	$[0.001, 0.99]$	λ^*	Gaussian	100	
20	300	$[100, 1000]$	$[0.001, 0.99]$	λ^*	linear	—	
21	300	$[100, 1000]$	$[0.001, 0.99]$	λ^*	Gaussian	—	

Table 2: Parameters used for each experiment, organized by Figure. L=left, C=center, R=right, U=upper, B=bottom