Robust Predictive Uncertainty and Double Descent in Contaminated Bayesian Random Features

Michele Caprio[1] **Katerina Papagiannouli**[2] **Siu Lun Chau**[3] **Sayan Mukherjee**[†4]

[1]The University of Manchester, UK
[2]University of Pisa, Italy
[3]Nanyang Technological University, Singapore
[4]Max Planck Institute for Mathematics in the Sciences, Germany

Abstract

We propose a robust Bayesian formulation of random feature (RF) regression that accounts explicitly for prior and likelihood misspecification via Huber-style contamination sets. Starting from the classical equivalence between ridge-regularized RF training and Bayesian inference with Gaussian priors and likelihoods, we replace the single prior and likelihood with ϵ- and η-contaminated credal sets, respectively, and perform inference using pessimistic generalized Bayesian updating. We derive explicit and tractable bounds for the resulting lower and upper posterior predictive densities. These bounds show that, when contamination is moderate, prior and likelihood ambiguity effectively acts as a direct contamination of the posterior predictive distribution, yielding uncertainty envelopes around the classical Gaussian predictive. We introduce an Imprecise Highest Density Region (IHDR) for robust predictive uncertainty quantification and show that it admits an efficient approximation via an adjusted Gaussian credible interval. We further obtain predictive variance bounds (under a mild truncation approximation for the upper bound) and prove that they preserve the leading-order proportional-growth asymptotics known for RF models. Together, these results establish a robustness theory for Bayesian random features: predictive uncertainty remains computationally tractable, inherits the classical double-descent phase structure, and is improved by explicit worst-case guarantees under bounded prior and likelihood misspecification.

1 INTRODUCTION

Random feature (RF) models are a classical and widely used mechanism to scale kernel methods to modern dataset sizes [Rahimi and Recht, 2007]. By replacing an implicit kernel with a finite collection of randomized nonlinear features, RF models turn nonlinear function learning into a linear problem in a randomized feature space, enabling fast training via least squares and ridge regression.

Beyond their computational appeal, RF models have become a central object in the study of high-dimensional learning. In proportional asymptotic regimes where the number of samples n, ambient dimension d, and number of random features N grow at comparable rates, prediction risk and predictive variance exhibit the now-classical double-descent phenomenon [Belkin et al., 2019]. A precise asymptotic theory has since been developed for ridgeless least squares and random feature regression [Mei and Montanari, 2022, Hastie et al., 2022], showing that risk and variance display singular behavior near the interpolation threshold $N \approx n$, corresponding to $\psi_1 = N/d$ approaching $\psi_2 = n/d$. These works reveal that double descent is not merely empirical, but a structural feature of high-dimensional models.

A complementary and increasingly popular viewpoint interprets ridge-regularized RF training as Bayesian inference. In particular, the RF ridge solution can be understood as a maximum a posteriori (MAP) estimator. In standard Gaussian constructions, it coincides with the mean of a posterior predictive distribution for new responses [Baek et al., 2023, Section 2.2]. This Bayesian lens naturally supports uncertainty quantification, provides a principled route to hyperparameter selection, and connects RF models to broader probabilistic and statistical learning paradigms.

However, the classical Bayesian RF formulation relies on modeling choices that are rarely exact in practice. The Gaussian prior on the random-feature coefficients and the Gaussian likelihood for the outputs are convenient and analytically tractable, but they can be misspecified due to heavy-tailed noise, label corruption, heteroskedasticity, or distribution shift between training and testing. Even small deviations from these assumptions may yield posterior predictives that are overly confident, brittle to outliers, or sensitive to

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:943–960.

prior choices. This motivates the question we study:

How do Bayesian RF conclusions change when we explicitly account for prior and likelihood misspecification?

Our approach: a Contamination Random Features model. We address this question through the lens of Bayesian sensitivity analysis [Berger, 1984] using Huber-style contamination sets [Walley, 1991, Huber and Ronchetti, 2009]. Concretely, instead of committing to a single prior and likelihood, we consider credal sets [Levi, 1980], i.e. convex and weak*-closed sets of probabilities. Specifically, we work with an ϵ-contaminated prior set and an η-contaminated likelihood set, obtained by mixing the baseline Gaussian specifications with arbitrary alternatives. This formalizes a principled "ambiguity budget" for both prior and data generating assumptions, and it also provides a way to model certain forms of distribution shift as long as the shift remains within the specified sets. This is reminiscent of the work in distributionally robust optimization with ambiguity sets [Dellaporta et al., 2024, Chen et al., 2026].

Choosing the contamination levels. The contamination levels ϵ and η should be interpreted as robustness or sensitivity parameters. When domain information is available, η may be chosen as an upper bound on the expected fraction of corrupted, outlying, or distribution-shifted observations, while ϵ encodes the analyst's tolerance to prior misspecification. In the absence of such information, we recommend reporting sensitivity curves over a grid of plausible values of (ϵ, η). When validation data are available, η can also be calibrated by choosing the smallest value for which the adjusted IHDR, or its Gaussian surrogate (see Section 3.1), attains the desired empirical predictive coverage on held-out data. Fully automatic selection of (ϵ, η), especially under arbitrary nonparametric contamination, requires additional modeling assumptions and is left for future work.

Remark 1 (Role of Bayesian Sensitivity Analysis)**.** *The purpose of Bayesian sensitivity analysis is to assess the stability of inferential conclusions under departures from baseline modeling assumptions. In our setting, given our "ambiguity budgets", we study how core predictive conclusions, such as predictive densities, uncertainty regions, and variance phase structure, change across this neighborhood. This is a robustness objective—yielding conservative lower/upper posterior predictives—rather than an attempt to propose a new predictor that uniformly improves test error.*

A subtle but important technical point is that ensuring weak* closedness—and hence the well-definedness of credal sets—may require working in the space of finitely additive probabilities [Walley, 1991]. While this generality can raise measure-theoretic complications, our analysis admits natural extremal contaminations (e.g., bounded spike densities concentrated near selected points), preserving an intuitive interpretation reminiscent of spike-and-slab behavior [Louzada et al., 2023].

Pessimistic generalized Bayes updating and predictive robustness. Within the credal sets framework, a well-established method of carrying out posterior inference is via a set of posteriors, equivalently summarized by its lower and upper envelopes, called lower and upper posteriors, respectively. We adopt the update rule of a pessimistic generalized Bayesian agent [Walley, 1991, Theorem 6.4.6], which yields conservative (worst-case) posterior conclusions under the specified contamination sets. Our primary object of interest is predictive: the lower and upper posterior predictive distributions for a new response $\tilde{y}$ at a test input $\tilde{\boldsymbol{x}}$.

Main results: tractable bounds for lower/upper posterior predictives. Our first contribution is an explicit characterization of lower and upper densities associated with contamination credal sets, and their role in posterior updating. Building on this, we derive simple and interpretable bounds for the lower and upper posterior predictive densities induced by simultaneous prior and likelihood contamination. Let ℓ_{pred} denote the classical RF posterior predictive density arising from the baseline Gaussian prior and likelihood (a one-dimensional normal distribution with mean given by the ridge RF predictor and variance given by the standard Bayesian RF expression [Baek et al., 2023]). We prove that the *lower* posterior predictive density $\underline{\ell}_{\mathrm{pred}}$ satisfies

$$\underline{\ell}_{\mathrm{pred}} \leq (1-\eta)\ell_{\mathrm{pred}},$$

and that the *upper* posterior predictive density $\overline{\ell}_{\mathrm{pred}}$ satisfies a complementary inequality. These bounds highlight a simple phenomenon: likelihood contamination acts as an adversarial "dilution" of the baseline predictive density (and, in the upper case, an additive worst-case term), while prior contamination enters through the normalization and extremal updates. Importantly, when ϵ and η are not too large, our derivations show that contaminating the prior and likelihood leads approximately to the same effect as directly contaminating the posterior predictive distribution, echoing recent observations in related ambiguity-set formulations [Dellaporta et al., 2024].

Uncertainty quantification via IHDRs. A second contribution is a predictive uncertainty notion tailored to our imprecise predictive goal: the *Imprecise Highest Density Region* (IHDR), defined as the smallest region that attains a target lower predictive probability mass. We show that, under the above predictive bounds, the IHDR can be approximated by a credible interval of the classical Gaussian posterior predictive distribution with an adjusted coverage level. This yields a practical, computable procedure for robust predictive intervals while preserving the analytic tractability of the RF Bayesian model.

Moment consequences and asymptotics. Finally, we connect these predictive density bounds to bounds on predictive dispersion. In particular, we upper bound the lower predic-

tive variance by a simple rescaling of the classical predictive variance, and, under a mild truncation approximation, we obtain a corresponding lower bound for the upper predictive variance. Because the contamination parameters are constants, these variance bounds inherit the high-dimensional asymptotic behavior characterized in the RF literature [Baek et al., 2023]. In other words, the Contamination RF model retains the same leading-order asymptotics as the baseline RF model, while providing a controlled robustness envelope around its predictions.

Contributions. In summary, our paper makes the following contributions:

- We introduce the *Contamination Random Features* model: a Bayesian RF formulation with ϵ-contaminated priors and η-contaminated likelihoods represented as credal sets.
- We characterize lower and upper densities for contamination sets and derive tractable, explicit bounds on the lower and upper posterior predictive densities.
- We define predictive *IHDRs* for imprecise posterior predictives and provide a practical approximation via adjusted Gaussian predictive credible intervals.
- We derive predictive variance bounds and show how these bounds align with known RF asymptotics in proportional-growth regimes.

Organization. The remainder of the paper is organized as follows. Section 2 introduces the RF model and its Bayesian interpretation. Section 3 presents our main predictive bounds and their implications for IHDR construction. Sections 4 and 5 derive predictive variance bounds and discuss asymptotics. Finally, in Section 6 we provide numerical experiments illustrating the predictive bounds and the preservation of double-descent structure under contamination. Section 7 concludes our work. Proofs and technical details are deferred to the appendix.

2 RANDOM FEATURE (RF) MODEL

Let us briefly give a background on the classical Random Features Model. Denote inputs $\boldsymbol{x}_i \in \mathbb{R}^d, i \in \{1, 2, \ldots, n\}$, as i.i.d. samples drawn from a uniform measure on a d-dimensional sphere with respect to the conventional Euclidean norm

$$\mathbb{S}^{d-1}(\sqrt{d}) := \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\| = \sqrt{d}\}.$$

Let outputs y_i be generated by the following model,

$$y_i = f_d(\boldsymbol{x}_i) + \gamma_i, \quad f_d(\boldsymbol{x}) = \beta_{d,0} + \langle \boldsymbol{x}, \boldsymbol{\beta}_d \rangle + f_d^{\mathrm{NL}}(\boldsymbol{x}).$$

The model is decomposed into a linear component, $\beta_{d,0} + \langle \boldsymbol{x}, \boldsymbol{\beta}_d \rangle$, and a nonlinear component, f_d^{NL}. The random errors γ_i's are assumed to be i.i.d. with mean zero, variance τ^2, and finite fourth moment. We allow $\tau^2 = 0$, in which case the generating model is *noiseless*.

We want to learn the optimal *Random Features Model* that best fits the training data. This is a class of functions

$$\mathcal{F} := \big\{ f : f(\boldsymbol{x}) = \sum_{j=1}^{N} a_j \sigma(\langle \boldsymbol{x}, \boldsymbol{\theta}_j \rangle / \sqrt{d}) \big\}, \tag{1}$$

which is dependent on N random features, $(\boldsymbol{\theta}_j)_{j=1}^N$, which are drawn i.i.d. from $\mathbb{S}^{d-1}$, and a nonlinear activation function, σ. The training objective solves a regularized least squares problem for the linear coefficients $\boldsymbol{a} \equiv (a_j)_{j=1}^N$,

$$\begin{aligned} \min_{\boldsymbol{a} \in \mathbb{R}^N} \Big\{ \sum_{i=1}^{n} \Big(y_i - \sum_{j=1}^{N} a_j \sigma \left(\langle \boldsymbol{x}_i, \theta_j \rangle / \sqrt{d} \right) \Big)^2 \\ + d \psi_{1,d} \psi_{2,d} \lambda \|\boldsymbol{a}\|^2 \Big\}, \end{aligned} \tag{2}$$

where $\psi_{1,d} = N/d$, $\psi_{2,d} = n/d$, and $\lambda > 0$. The optimal weights, $\widehat{\boldsymbol{a}} \equiv \widehat{\boldsymbol{a}}(\lambda)$, determine an optimal ridge predictor denoted by $\widehat{f} \equiv f(\cdot; \widehat{\boldsymbol{a}}(\lambda))$.[1] The dependence of the trained predictor on the dataset $(\boldsymbol{X}, \boldsymbol{y})$ and features $\boldsymbol{\Theta}$ are suppressed in notation.

The objective function (2) of a classical Random Features (RF) model can be interpreted as a MAP estimation problem for an equivalent Bayesian model.[2] As we shall see more in detail later, the goal of this paper is to provide the Bayesian sensitivity analysis [Berger, 1984] version of the Bayesian RF model, via contamination sets [Huber and Ronchetti, 2009]. We call it the *Contamination RF Model*.

Formally, we adopt a d-dependent weight prior distribution, denoted $p(\boldsymbol{a})$, and also a normal likelihood, denoted $p(\boldsymbol{y} | \boldsymbol{X}, \boldsymbol{\Theta}, \boldsymbol{a})$, centered around a function in the class (1) with variance ϕ^{-1},

$$\boldsymbol{a} \sim \mathcal{N}\left(0, \phi^{-1} \frac{\psi_{1,d} \psi_{2,d} \lambda}{d} \mathbf{I}_N\right), \tag{3}$$

$$\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\Theta}, \boldsymbol{a} \sim \mathcal{N}\left(\sigma(\boldsymbol{X}\boldsymbol{\Theta}/\sqrt{d})\boldsymbol{a}, \phi^{-1} \mathbf{I}_n\right). \tag{4}$$

3 BOUNDING THE POSTERIOR PREDICTIVE DENSITY

We now turn to the central question of the paper: *how does predictive inference in Bayesian random feature models change under explicit prior and likelihood misspecification?* To address this, we replace the single Gaussian prior and likelihood by contamination credal sets and analyze the resulting lower and upper posterior predictive. To face such a possibility, we specify an ϵ-contaminated prior credal set

[1] Given a new input feature $\tilde{\boldsymbol{x}}$, the ridge regularized predictor $\widehat{f}(\tilde{\boldsymbol{x}}) \equiv f(\tilde{\boldsymbol{x}}; \widehat{\boldsymbol{a}})$ for the unknown output $\tilde{y}$ is given in Baek et al. [2023, Equation (5)].

[2] Here, a MAP has to be understood with respect to a posterior predictive distribution.

and an η-contaminated likelihood credal set. Formally, we let $P^{\mathcal{N}}$ denote the (countably additive) probability measure whose pdf is given by the Normal in (3)—which, for notational convenience, we denote by $p^{\mathcal{N}}$. Similarly, let $L^{\mathcal{N}} \equiv P(\cdot \mid \boldsymbol{X}, \boldsymbol{\Theta}, \boldsymbol{a})$ denote the probability measure whose pdf is given by the Normal in (4)—which, for notational convenience, we denote by $\ell^{\mathcal{N}}_{\boldsymbol{a}} \equiv p(\cdot|\boldsymbol{X}, \boldsymbol{\Theta}, \boldsymbol{a})$. In other words, given a sigma-finite dominating measure ν such as the Lebesgue one, we let $\mathrm{d}P^{\mathcal{N}}/\mathrm{d}\nu = p^{\mathcal{N}}$ and $\mathrm{d}L^{\mathcal{N}}/\mathrm{d}\nu = \ell^{\mathcal{N}}_{\boldsymbol{a}}$. Here, the ratios represent Radon-Nikodym derivatives. In the remainder of the paper, ν is left implicit for ease of exposition.

Pick any $\epsilon, \eta \in [0, 1]$ and denote by $\Delta^{\text{fa}}_{\mathbb{R}^N}$ and $\Delta^{\text{fa}}_{\mathbb{R}^n}$ the space of finitely additive probability measures on $\mathbb{R}^N$ and $\mathbb{R}^n$, respectively. Then, the prior and likelihood contamination credal sets are the following,

$$\mathcal{P}_\epsilon \coloneqq \{P \in \Delta^{\text{fa}}_{\mathbb{R}^N} : P = (1-\epsilon)P^{\mathcal{N}} + \epsilon Q, \forall Q \in \Delta^{\text{fa}}_{\mathbb{R}^N}\},$$
$$\mathcal{L}_\eta \coloneqq \{L \in \Delta^{\text{fa}}_{\mathbb{R}^n} : L = (1-\eta)L^{\mathcal{N}} + \eta S, \forall S \in \Delta^{\text{fa}}_{\mathbb{R}^n}\}.$$

Intuition for the contamination sets. The main intuition is simple: the ϵ-contaminated prior keeps a fraction $(1-\epsilon)$ of the baseline Gaussian prior $P^{\mathcal{N}}$, while allowing the remaining ϵ fraction to be placed adversarially. Therefore, for any proper measurable event A, the smallest prior probability assigned to A is $(1-\epsilon)P^{\mathcal{N}}(A)$, while the largest is $(1-\epsilon)P^{\mathcal{N}}(A) + \epsilon$. The likelihood contamination set has the same interpretation, with η controlling the amount of likelihood misspecification. The finitely additive formulation below is used to make these credal sets mathematically closed and well-defined; the predictive bounds derived later depend only on the lower and upper envelope formulas.

We need to consider finitely additive measures because, if we only considered countably additive ones, then the sets may be empty or not weak* closed, and hence not well-defined credal sets [Caprio, 2025, Section 3]. It is easy to see that both $\mathcal{P}_\epsilon$ and $\mathcal{L}_\eta$ are convex [Marinacci and Montrucchio, 2004]. Recall that the lower envelope of a credal set is called *lower probability*. It is the smallest value that the elements of the credal set can take on; in our case, we have $\underline{P} \coloneqq \inf_{P \in \mathcal{P}_\epsilon} P$ and $\underline{L} \coloneqq \inf_{L \in \mathcal{L}_\eta} L$. The upper envelope is called *upper probability* (denoted by $\overline{P}$ and $\overline{L}$, respectively), and is defined analogously, with sup in place of inf. Notice that lower and upper probabilities are conjugate, that is, for all $A \in \mathcal{B}(\mathbb{R}^N)$, $\underline{P}(A) = 1 - \overline{P}(A^c)$, where $\mathcal{B}(\mathbb{R}^N)$ denotes the Borel sigma-algebra of $\mathbb{R}^N$, and similarly for the lower and upper likelihood. The following is an important property of contamination sets, proved in Caprio [2025, Lemma 5].[3]

Lemma 1 (Characterizing the Contamination Sets)**.** *The following are true,*

$$\underline{P}(A) = \begin{cases} (1-\epsilon)P^{\mathcal{N}}(A) & A \in \mathcal{B}(\mathbb{R}^N) \setminus \{\mathbb{R}^N\} \\ 1 & A = \mathbb{R}^N \end{cases}$$

and

$$\overline{P}(A) = \begin{cases} (1-\epsilon)P^{\mathcal{N}}(A) + \epsilon & A \in \mathcal{B}(\mathbb{R}^N) \setminus \{\emptyset\} \\ 0 & A = \emptyset \end{cases}.$$

Similar characterization holds for $\overline{L}$ and $\underline{L}$. In addition,

$$\mathcal{P}_\epsilon = \{P \in \Delta^{fa}_{\mathbb{R}^N} : P(A) \geq \underline{P}(A), \forall A \in \mathcal{B}(\mathbb{R}^N)\} =: \mathcal{M}(\underline{P}),$$
$$\mathcal{L}_\eta = \{L \in \Delta^{fa}_{\mathbb{R}^n} : L(B) \geq \underline{L}(B), \forall B \in \mathcal{B}(\mathbb{R}^n)\} =: \mathcal{M}(\underline{L}). \tag{5}$$

The sets $\mathcal{M}(\underline{P})$ and $\mathcal{M}(\underline{L})$ are called the *cores* of $\underline{P}$ and $\underline{L}$, respectively. They are defined as the collection of all (finitely additive) probabilities that set-wise dominate the lower probabilities. Thanks to Equation (5), we know that lower probabilities are sufficient to characterize the entire credal set—they are *compatible* with the credal sets [Gong and Meng, 2021].

To avoid unnecessary complications, we perpetrate abuse of notation and write $\underline{P}(A) = (1-\epsilon)P^{\mathcal{N}}(A)$, $\overline{P}(A) = (1-\epsilon)P^{\mathcal{N}}(A) + \epsilon$, for proper measurable events A, leaving implicit the full space and empty set conventions stated in Lemma 1. We proceed similarly for $\overline{L}$; the interested reader can find more details around this choice in Caprio [2025, Section 3 and Appendix A].

The fundamental objects in the contamination model are the lower and upper probabilities, not unique lower or upper densities. When we restrict attention to an absolutely continuous subclass of contaminations, it is useful to introduce density-level representations associated with these lower and upper probabilities. We refer to these as lower- and upper-envelope densities. This terminology should not be understood as implying uniqueness, nor as implying that the envelopes are themselves additive probability measures with ordinary probability densities. Lemma 2 records these density-level representations as a technical device for the calculations below.

Lemma 2 (Density-level representations of lower and upper probabilities)**.** *The lower density $\underline{p}$ associated with lower probability $\underline{P}$ is such that, for all $A \in \mathcal{B}(\mathbb{R}^N) \setminus \{\mathbb{R}^N\}$, there exists a probability measure $Q_{\star A} \in \Delta^{fa}_{\mathbb{R}^N}$ with $Q_{\star A}(A) = 0$ and pdf $q_{\star A}$ such that*

$$\underline{P}(A) = \int_A \underbrace{[(1-\epsilon)p^{\mathcal{N}}(\boldsymbol{a}) + \epsilon q_{\star A}(\boldsymbol{a})]}_{=:\underline{p}(\boldsymbol{a})} d\boldsymbol{a}.$$

The upper density $\overline{p}$ associated with upper probability $\overline{P}$ is such that, for all $A \in \mathcal{B}(\mathbb{R}^N) \setminus \{\emptyset\}$, there exists a probability measure $Q^\star_A \in \Delta^{fa}_{\mathbb{R}^N}$ with $Q^\star_A(A) = 1$ and pdf $q^\star_A$ such that

$$\overline{P}(A) = \int_A \underbrace{[(1-\epsilon)p^{\mathcal{N}}(\boldsymbol{a}) + \epsilon q^\star_A(\boldsymbol{a})]}_{=:\overline{p}(\boldsymbol{a})} d\boldsymbol{a}.$$

[3] See also Wasserman and Kadane [1990], Walley [1991].

The reader may view Lemma 2 as a formal device for representing lower and upper envelope probabilities at the density level. The subsequent predictive results use only envelope bounds and do not depend on choosing a unique extremal density.

Three remarks are in order. First, the choices of $Q_{\star A}$ and $Q_A^\star$ depend on the set A, and so do the associated quantities $\underline{p}(\boldsymbol{a})$ and $\overline{p}(\boldsymbol{a})$. Thus, these quantities should be understood as density-level representations associated with the lower and upper probabilities, not as unique probability densities selected from the credal set, nor as implying that the envelopes themselves are additive probability measures with ordinary densities. At the set-function level, the fundamental objects remain $\underline{P}(A) = (1-\epsilon)P^{\mathcal{N}}(A), \qquad \overline{P}(A) = (1-\epsilon)P^{\mathcal{N}}(A) + \epsilon$, for proper measurable events A, together with the full-space and empty-set conventions stated in Lemma 1. Within the absolutely continuous subclass, the lower probability admits the subdensity representation $(1-\epsilon)p^{\mathcal{N}}$ on proper measurable events. In general, no finite A-independent upper-envelope density representation exists without further restrictions on the contaminating densities; for instance, if the contaminating densities are M-bounded a.e., for some $M > 0$, then $(1-\epsilon)p^{\mathcal{N}} + \epsilon M$ is a valid pointwise majorant.

Second, since $Q_{\star A}, Q_A^\star \in \Delta^{\mathrm{fa}}_{\mathbb{R}^N}$, they may be finitely additive but not countably additive. In that case, the concept of a pdf is only well-defined in the sense of a Banach limit, which may not be unique, or as the pdf of a Loeb measure, which requires nonstandard analysis. In the present setting, we avoid these complications: at the set-function level, it is convenient to think of the extremizers heuristically as Dirac masses.

Third, whenever we write lower/upper *densities*, we implicitly restrict attention to an absolutely continuous subclass of contaminations, so that all density-valued quantities and Lebesgue integrals appearing below are well-defined in the ordinary sense. In particular, whenever we write $\delta_{\mathbf{y}}$ or $\delta_{\tilde{y}}$ inside expressions for lower/upper densities, such as $\underline{\ell}_{\boldsymbol{a}}(\mathbf{y})$ or $\overline{\ell}_{\boldsymbol{a}}(\mathbf{y})$, we mean a bounded spike density $s_{\mathbf{y}}$ with respect to Lebesgue measure, satisfying $s_{\mathbf{y}} \geq 0$, $\int s_{\mathbf{y}}(\mathbf{z})d\mathbf{z} = 1$, and $\|s_{\mathbf{y}}\|_\infty < \infty$, for instance the uniform density on a small hyper-rectangle centered at $\mathbf{y}$. We keep the notation $\delta_{\mathbf{y}}$ for readability. Similarly, whenever we write $\delta_{\boldsymbol{a}_\star}$ and $\delta_{\boldsymbol{a}^\star}$ for lower/upper prior densities, we mean bounded spike densities concentrated in small neighborhoods of $\boldsymbol{a}_\star$ and $\boldsymbol{a}^\star$, respectively.

Remark 2 (On Envelope Densities and One-Sided Bounds)**.** *In what follows, $\underline{\ell}_{\mathrm{pred}}$ and $\overline{\ell}_{\mathrm{pred}}$ denote density-level representations associated with the lower and upper posterior predictive probabilities, within the absolutely continuous subclass described above.*

The bounds we derive are one-sided comparison bounds: (i) an upper bound *for the lower posterior predictive envelope density $\underline{\ell}_{\mathrm{pred}}$, and (ii) a* lower bound *for the upper posterior predictive envelope density $\overline{\ell}_{\mathrm{pred}}$. They should not be read as a full two-sided cautious characterization of the posterior predictive credal set. Rather, they show how prior and likelihood contamination compare to direct contamination of the classical Gaussian posterior predictive.*

Consequently, we do not *require the existence of a finite, A-independent pointwise upper-envelope density for the contaminating prior densities. The derivations use only the lower/upper probability envelopes and their density-level representations within the absolutely continuous subclass, and therefore do not depend on the existence of such a pointwise majorant.*

Recall now that the optimal ridge weights $\hat{\boldsymbol{a}}$ that solve (2) coincide with the posterior mean of the weight vector $\boldsymbol{a}$ under prior (3) and likelihood (4). Consequently, given a new input feature $\tilde{\boldsymbol{x}}$, the posterior predictive distribution of $\tilde{y}$ induced by (3) and (4) has mean $\hat{f}(\tilde{\boldsymbol{x}})$. Hence, we are now interested in deriving the lower posterior predictive density. Notice that we assume that the learner is a *pessimistic generalized Bayesian* agent [Walley, 1991, Theorem 6.4.6], that is, they update their beliefs in the form of an updated lower probability as

$$\underline{P}(\boldsymbol{a} \in A \mid \boldsymbol{X}, \boldsymbol{\Theta}, \boldsymbol{y} \in B) = \frac{\underline{P}(A)\underline{P}(B \mid \boldsymbol{X}, \boldsymbol{\Theta}, A)}{\overline{P}(B \mid \boldsymbol{X}, \boldsymbol{\Theta})}, \quad (6)$$

with $A \in \mathcal{B}(\mathbb{R}^N)$, $B \in \mathcal{B}(\mathbb{R}^n)$. For other types of updates, we refer the interested reader to Walley [1991], Gong and Meng [2021], Caprio and Seidenfeld [2023]; here, we limit ourselves to pointing out that pessimistic generalized Bayesian updating is a lower bound for both Walley's classical generalized Bayesian updating [Walley, 1991] and the geometric updating rule [Suppes and Zanotti, 1977].[4]

Let $\ell_{\mathrm{pred}} \equiv p(\cdot \mid \tilde{\boldsymbol{x}}, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\Theta})$ denote the density associated with the posterior predictive distribution "induced" by the Bayesian RF model encoded in (3) and (4), and a new input $\tilde{\boldsymbol{x}}$. It is a one-dimensional Normal $\mathcal{N}(\hat{f}(\tilde{\boldsymbol{x}}), s^2(\tilde{\boldsymbol{x}}))$ whose parameters are given in Baek et al. [2023, Section 2.2].

Theorem 3 (Bounding the Lower Posterior Predictive Density)**.** *Let ℓ_{pred} denote the pdf of the one-dimensional posterior predictive Normal distribution $\mathcal{N}(\hat{f}(\tilde{\boldsymbol{x}}), s^2(\tilde{\boldsymbol{x}}))$. Then, the following is true for all $\tilde{y} \in \mathbb{R}$,*

$$\underline{\ell}_{pred}(\tilde{y}) \leq (1-\eta)\ell_{pred}(\tilde{y}),$$

where $\underline{\ell}_{pred}$ denotes the lower posterior predictive density.

In the following corollary, we bound the upper posterior predictive density.

[4] We will study its coherence [Walley, 1991, Section 2.5] in future work.

Corollary 3.1 (Bounding the Upper Posterior Predictive Density)**.** *The following is true for all* $\tilde{y} \in \mathbb{R}$*,*

$$\overline{\ell}_{pred}(\tilde{y}) \geq (1-\eta)\ell_{pred}(\tilde{y}) + \eta u(\tilde{y}),$$

where $\overline{\ell}_{pred}$ *denotes the upper posterior predictive density and* u *is any probability density on* $\mathbb{R}$ *(for convenience, one may take* u *bounded).*

Interpretation of the one-sided bounds. The inequalities in Theorem 3 and Corollary 3.1 should be interpreted as one-sided comparison bounds for the density-level representations of the posterior predictive envelopes. They show that, in the sense described above, prior and likelihood contamination lead to behavior comparable to a direct contamination of the classical Gaussian posterior predictive. They are not intended to give a full two-sided cautious characterization of the posterior predictive credal set. In particular, bounds in the opposite directions—a lower bound for the lower posterior predictive envelope or an upper bound for the upper posterior predictive envelope—would require additional restrictions on the contaminating class, such as boundedness, support, or moment conditions. Without such restrictions, a finite A-independent pointwise upper-envelope density, and hence a finite uniform upper variance bound, need not exist.

As we can see from Theorem 3 and Corollary 3.1, if the prior and likelihood contaminations are not too extreme, contaminating prior and likelihood leads approximately to the same result as directly contaminating the posterior predictive distribution. This is conceptually related to the Bayesian ambiguity set perspective of Dellaporta et al. [2024], where posterior-informed ambiguity at the model level is used for downstream distributionally robust optimization. Our focus is different: rather than optimizing a worst case decision objective, we study robust posterior predictive inference for Bayesian random features, deriving explicit lower/upper predictive density bounds, IHDR approximations, and variance envelopes that preserve the double descent structure in proportional growth regimes.

3.1 APPROXIMATING THE IMPRECISE HIGHEST DENSITY REGION

The lower posterior predictive density $\underline{\ell}_{\text{pred}}$ can be used to define the lower posterior predictive distribution as $\underline{P}(\tilde{y} \in B \mid \tilde{x}, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\Theta}) := \int_B \underline{\ell}_{\text{pred}}(\tilde{y})\mathrm{d}\tilde{y}$, for all $B \in \mathcal{B}(\mathbb{R}) \setminus \{\mathbb{R}\}$, and $\underline{P}(\tilde{y} \in \mathbb{R} \mid \tilde{x}, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\Theta}) = 1$. Just as in the proof of Theorem 3, it can be routinely checked that $\underline{P}_{\text{pred}} \equiv \underline{P}(\cdot \mid \tilde{x}, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\Theta})$ is a well-defined lower probability. In turn, we can use such a lower probability to derive a predictive Imprecise Highest Density Region [Coolen, 1992].

Definition 4 (Imprecise Highest Density Region—IHDR)**.** *Let* α *be any value in* $[0,1]$*. Then, the set* $IR_\alpha[\mathcal{M}(\underline{P}_{pred})] \in \mathcal{B}(\mathbb{R})$ *is called a* $(1-\alpha)$*-Imprecise Highest Density Region (IHDR) if*

1. $\underline{P}_{pred}\left(\{\tilde{y} \in IR_\alpha[\mathcal{M}(\underline{P}_{pred})]\}\right) = 1-\alpha$*;*
2. $\int_{IR_\alpha[\mathcal{M}(\underline{P}_{pred})]} d\tilde{y}$ *is a minimum.*

Notice that Condition 2 is needed so that $\text{IR}_\alpha[\mathcal{M}(\underline{P}_{\text{pred}})]$ is the smallest possible subset of $\mathbb{R}$, which still satisfies Condition 1. By the definition of lower probability, Definition 4 implies that $P_{\text{pred}}(\{\tilde{y} \in \text{IR}_\alpha[\mathcal{M}(\underline{P}_{\text{pred}})]\}) \geq 1-\alpha$, for all $P \in \mathcal{M}(\underline{P}_{\text{pred}})$. Here lies the appeal of the IHDR concept. The following shows that it easy to approximate the IHDR.

Proposition 5 (Approximating the IHDR)**.** *Pick any* $\alpha \in [0,1]$*. Let* $\beta := \min\{1, \frac{1-\alpha}{1-\eta}\}$*. If* $IR_\alpha[\mathcal{M}(\underline{P}_{pred})]$ *exists, then*

$$\int_{R_\beta[P_{pred}]} \mathrm{d}\tilde{y} \leq \int_{IR_\alpha[\mathcal{M}(\underline{P}_{pred})]} \mathrm{d}\tilde{y},$$

where $R_\beta[P_{pred}]$ *denotes the* β*-level credible interval of the posterior predictive distribution* $\ell_{pred} = \mathcal{N}(\hat{f}(\tilde{\boldsymbol{x}}), s^2(\tilde{\boldsymbol{x}}))$*. In particular,* $R_\beta[P_{pred}]$ *is a computable Gaussian surrogate for the IHDR.*

Empirical choice of η for IHDR calibration. The adjusted IHDR approximation depends on the likelihood-contamination level η. When labeled calibration data of the form $(x_i^{\text{cal}}, y_i^{\text{cal}})$ are available, η can be chosen empirically. For a target coverage level $1-\alpha$, one may evaluate the Gaussian surrogate intervals $C_\eta(x) := R_{\beta_\eta}[P_{\text{pred}}(\cdot \mid x)]$, $\beta \equiv \beta_\eta := \min\left\{1, \frac{1-\alpha}{1-\eta}\right\}$, over a grid of values $\eta \in \mathcal{G}$, and select the smallest η such that

$$\frac{1}{m}\sum_{i=1}^{m} \mathbf{1}\{y_i^{\text{cal}} \in C_\eta(x_i^{\text{cal}})\} \geq 1-\alpha,$$

where $\mathbf{1}\{\cdot\}$ denotes the indicator function. Among values satisfying this coverage constraint, one may further minimize average interval width or a proper interval score. Under arbitrary nonparametric contamination, η is not identifiable from unlabeled data alone without additional assumptions on the contamination mechanism. If no labeled calibration data or prior knowledge about misspecification are available, then η should be treated as a sensitivity parameter rather than as an identifiable quantity; in that case, the recommended practice is to report predictive envelopes over a range of plausible contamination levels.

4 BOUNDING THE VARIANCE

In this section, we study bounds for the lower and upper predictive variances of the contaminated RF model. As a byproduct, we are able to state a result on the double descent behavior of our robust model.

Inspired by Walley [1991, Appendix G.1], we define the *lower predictive variance* as

$$\mathbb{V}_{\underline{\ell}_{\text{pred}}}(\tilde{Y}) \coloneqq \min_{\zeta\in\mathbb{R}} \mathbb{E}_{\underline{\ell}_{\text{pred}}}[(\tilde{Y}-\zeta)^2]$$
$$= \min_{\zeta\in\mathbb{R}} \int_{\mathbb{R}} [\tilde{y}-\zeta]^2 \underline{\ell}_{\text{pred}}(\tilde{y})\mathrm{d}\tilde{y}.$$

Then, we have the following.

Proposition 6 (Bounding the Lower Predictive Variance)**.** *Let $\tilde{Y}$ denote the predictive quantity at $\tilde{x}$. Under the classical posterior predictive, $\tilde{Y}$ has density $\ell_{pred} = \mathcal{N}(\hat{f}(\tilde{x}), s^2(\tilde{x}))$, whereas under the lower posterior predictive it has density $\underline{\ell}_{pred}$. Then, the following is true,*

$$\mathbb{V}_{\underline{\ell}_{pred}}(\tilde{Y}) \leq (1-\eta)\mathbb{V}_{\ell_{pred}}(\tilde{Y}).$$

Keeping the same notation, under an additional assumption, we can also bound the upper predictive variance, whose definition is once again inspired by Walley [1991, Appendix G.1], $\mathbb{V}_{\overline{\ell}_{\text{pred}}}(\tilde{Y}) \coloneqq \min_{\zeta\in\mathbb{R}} \mathbb{E}_{\overline{\ell}_{\text{pred}}}[(\tilde{Y}-\zeta)^2]$.

Proposition 7 (Bounding the Upper Predictive Variance)**.** *Suppose we approximate the classical RF posterior predictive distribution $\ell_{pred} = \mathcal{N}(\hat{f}(\tilde{\boldsymbol{x}}), s^2(\tilde{\boldsymbol{x}}))$ with a truncated version $\ell'_{pred} = \mathcal{N}_{[a,b]}(\hat{f}(\tilde{\boldsymbol{x}}), s^2(\tilde{\boldsymbol{x}}))$, where $[a,b] \subset \mathbb{R}$, $a < b$. Let*

$$m \coloneqq \int_a^b \ell_{pred}(\tilde{y})\, d\tilde{y}, \qquad u_{[a,b]}(\tilde{y}) \coloneqq \frac{1}{b-a}\mathbf{1}_{[a,b]}(\tilde{y}).$$

Then,

$$\mathbb{V}_{\overline{\ell}_{pred}}(\tilde{Y}) \geq (1-\eta)m\,\mathbb{V}_{\ell'_{pred}}(\tilde{Y}) + \eta\frac{(b-a)^2}{12}.$$

In particular, if $[a,b]$ captures most of the Gaussian mass of ℓ_{pred}, so that $m \approx 1$, then

$$\mathbb{V}_{\overline{\ell}_{pred}}(\tilde{Y}) \gtrsim (1-\eta)\mathbb{V}_{\ell'_{pred}}(\tilde{Y}) + \eta\frac{(b-a)^2}{12}.$$

We sum up our finding on the variance in the following.

Lemma 8 (Predictive Variance Bounds)**.** *If $\ell_{pred} \approx \ell'_{pred}$, $m \approx 1$, and $\frac{(b-a)^2}{12} \geq \mathbb{V}_{\ell_{pred}}(\tilde{Y})$, then it holds that*

$$\begin{aligned}\mathbb{V}_{\underline{\ell}_{pred}}(\tilde{Y}) &\leq (1-\eta)\mathbb{V}_{\ell_{pred}}(\tilde{Y}) \leq \mathbb{V}_{\ell_{pred}}(\tilde{Y}) \\ &\leq (1-\eta)\mathbb{V}_{\ell_{pred}}(\tilde{Y}) + \eta\frac{(b-a)^2}{12} \\ &\approx (1-\eta)\mathbb{V}_{\ell'_{pred}}(\tilde{Y}) + \eta\frac{(b-a)^2}{12} \lesssim \mathbb{V}_{\overline{\ell}_{pred}}(\tilde{Y}).\end{aligned} \tag{7}$$

We note in passing that the conditions in Lemma 8 are always satisfied, because it is the user that selects a and b, and so they can pick them to meet the requirements $m \approx 1$ and $\frac{(b-a)^2}{12} \geq \mathbb{V}_{\ell_{\text{pred}}}(\tilde{Y})$. For example, this holds if we use the classic heuristic of $a = \hat{f}(\tilde{\mathbf{x}}) - 3s(\tilde{\mathbf{x}})$ and $b = \hat{f}(\tilde{\mathbf{x}}) + 3s(\tilde{\mathbf{x}})$.

In addition, we use the truncation in Proposition 7 because, without it, the function $\tilde{y} \mapsto \varphi(\tilde{y}) \coloneqq (\tilde{y}-\zeta)^2$ need not be (Choquet) integrable with respect to the upper probability associated with the bound for the upper predictive density found in Corollary 3.1 [Troffaes and de Cooman, 2014, Appendix C].

Remark 3 (On the Choice of the Truncation Interval)**.** *The bound on the upper predictive variance in Proposition 7 depends on the truncation interval $[a,b]$, through the terms m and $(b-a)^2/12$. If the interval is too narrow, the truncated predictive density ℓ'_{pred} deviates substantially from the original Gaussian predictive density ℓ_{pred}. If the interval is too wide, the term $(b-a)^2/12$ becomes unnecessarily large, leading to overly conservative uncertainty estimates.*

On the one hand, to ensure that the truncated predictive density ℓ'_{pred} remains close to the Gaussian predictive density $\ell_{\text{pred}} = \mathcal{N}\Big(\hat{f}(\tilde{\mathbf{x}}), s^2(\tilde{\mathbf{x}})\Big)$, the interval $[a,b]$ should contain most of the Gaussian mass. For the symmetric choice

$$a = \hat{f}(\tilde{\mathbf{x}}) - ks(\tilde{\mathbf{x}}), \qquad b = \hat{f}(\tilde{\mathbf{x}}) + ks(\tilde{\mathbf{x}}),$$

the retained mass is $m_k = \int_a^b \ell_{\text{pred}}(\tilde{y})\, d\tilde{y} = 2\Phi(k) - 1$, where Φ denotes the standard normal distribution function. Hence the truncation error is explicitly $1 - m_k = 2\{1 - \Phi(k)\}$. For example, $k = 2$ gives $m_k \approx 0.9545$, while $k = 3$ gives $m_k \approx 0.9973$. On the other hand, increasing k enlarges $\frac{(b-a)^2}{12} = \frac{k^2}{3}s^2(\tilde{\mathbf{x}})$, thereby enlarging the uniform-contamination contribution to the upper variance envelope.

In the proportional asymptotic regime, a natural scaling choice is $k = O(1)$, independent of dimension, ensuring that the truncation width remains proportional to the predictive standard deviation $s(\tilde{\mathbf{x}})$. Determining an optimal or data-adaptive choice of $[a,b]$ – for instance, via calibration criteria or decision-theoretic objectives – is an interesting direction for future work.

5 VARIANCE AND THE DOUBLE DESCENT PHENOMENON

In this section, we work in the proportional growth random features regime, where

$$d, n, N \to \infty \quad \text{with} \quad \psi_1 \coloneqq {}^{N}\!/_{d}, \quad \psi_2 \coloneqq {}^{n}\!/_{d}$$

fixed, and consider ridge-regularized random features regression with penalty $\lambda > 0$.

Proof roadmap and interpretation. The double-descent preservation result below should be understood as the final step of the robustness analysis developed in Sections 3 and

4. The technically substantive part is the derivation of the lower and upper posterior predictive density bounds and their translation into the variance envelope of Lemma 8. Once this envelope has been obtained, the preservation of the interpolation peak follows from the fact that the contamination bounds act as positive rescalings, or under the truncation choice below proportional transformations, of the baseline predictive variance. Thus, the result does not claim that affine invariance is technically difficult; rather, it shows that the robust contamination model reduces to a form that preserves the known double-descent phase structure of random-features regression.

Let $\ell_{\mathrm{pred}} = \mathcal{N}(\hat{f}(\tilde{x}), s^2(\tilde{x}))$ denote the classical Bayesian RF posterior predictive distribution induced by a Gaussian prior and Gaussian likelihood, and define the baseline predictive variance $V_{\mathrm{base}}(\psi_1; \lambda) \coloneqq \mathbb{V}_{\ell_{\mathrm{pred}}}(\tilde{Y})$.

To highlight the dependence of the bounds for the variance on η, let $\underline{V}(\psi_1; \lambda, \eta) \coloneqq \mathbb{V}_{\underline{\ell}_{\mathrm{pred}}}(\tilde{Y})$, $\overline{V}(\psi_1; \lambda, \eta) \coloneqq \mathbb{V}_{\overline{\ell}_{\mathrm{pred}}}(\tilde{Y})$. We now study formally the double descent phenomenon in the contamination credal set case.

Corollary 8.1 (Persistence of Double Descent Under Contamination)**.** *Fix $\psi_2 = n/d$ and $\lambda \geq 0$. Assume throughout that $\eta \in [0, 1)$, so that $1 - \eta > 0$. Suppose that the baseline predictive variance $V_{\mathrm{base}}(\psi_1; \lambda)$ exhibits double descent as a function of ψ_1, with a nondegenerate local maximizer at $\psi_1^\star$. Then the bounding functions appearing in Lemma 8 preserve the location of this peak:*

1. *The function*

$$g_-(\psi_1) \coloneqq (1 - \eta) V_{\mathrm{base}}(\psi_1; \lambda)$$

has the same maximizer(s) as $V_{\mathrm{base}}(\psi_1; \lambda)$.

2. *Under the truncation approximation with*

$$a(\psi_1) = \hat{f}(\tilde{\mathbf{x}}) + \alpha_0\, s(\tilde{\mathbf{x}}), \quad b(\psi_1) = \hat{f}(\tilde{\mathbf{x}}) + \beta_0\, s(\tilde{\mathbf{x}}),$$

for fixed constants $\alpha_0 < \beta_0$ independent of ψ_1, the function

$$g_+(\psi_1) \coloneqq (1 - \eta) V'_{\mathrm{base}}(\psi_1; \lambda) + \eta \frac{(b - a)^2}{12},$$

where $V'_{\mathrm{base}}(\psi_1; \lambda)$ is the predictive variance under ℓ'_{pred}, has the same maximizer(s) as $V_{\mathrm{base}}(\psi_1; \lambda)$.

In particular, the bounding functions form an envelope around the double-descent peak $\psi_1^\star$.

6 NUMERICAL EXPERIMENTS

In this section, we investigate the mechanism underlying the double-descent phenomenon and its robustness to likelihood misspecification.

Specifically, we decompose the test error into bias and variance terms, and then study how likelihood contamination induces uncertainty envelopes around the variance. The purpose of these experiments is deliberately targeted: rather than providing a broad empirical benchmark, we use simulations to illustrate the two theoretical mechanisms established above, namely variance-driven double descent and the preservation of the interpolation peak under contamination-induced uncertainty envelopes. We emphasize that the proposed framework is not intended to improve point-prediction accuracy directly; its practical benefit is robust predictive uncertainty quantification, namely producing conservative predictive envelopes and adjusted uncertainty regions under misspecification.

Bias-variance decomposition. For each feature-to-dimension ratio $\psi_1 = N/d$, we estimate the decomposition

$$\mathbb{E}\big[(\hat{f}(x) - f^*(x))^2\big] = \underbrace{\big(\mathbb{E}[\hat{f}(x)] - f^*(x)\big)^2}_{\text{Bias}^2} + \underbrace{\mathbb{E}\big[(\hat{f}(x) - \mathbb{E}[\hat{f}(x)])^2\big]}_{\text{Variance}},$$

where the expectation is taken over independent draws of the training data and random features. Figure 2.(left) in Appendix A shows that the double descent peak is entirely driven by the variance term, while the bias varies smoothly with ψ_1. This confirms that the interpolation instability in random-feature regression is a variance-dominated phenomenon, especially in the ridgeless case, consistent with existing random matrix analyses.

Contamination-induced variance envelopes. We then visualize the effect of likelihood contamination on predictive uncertainty. Lemma 8 shows that the lower and upper predictive variances under contamination level η satisfy affine bounds of the form

$$\underline{V}(\psi_1; \lambda, \eta) \leq (1 - \eta) V_{\mathrm{base}}(\psi_1; \lambda) \leq \overline{V}(\psi_1; \lambda, \eta),$$

where $V_{\mathrm{base}}(\psi_1; \lambda)$ denotes the baseline predictive variance. Corollary 8.1 further implies that such affine transformations preserve the location of any double descent peak in ψ_1.

Figure 2.(right) in Appendix A illustrates contamination-induced variance envelopes constructed via affine transformations of the baseline predictive variance. The lower envelope corresponds to the scaling $(1 - \eta)V_{\mathrm{base}}$, while the upper envelope is constructed from the truncated variance bound of Proposition 7, namely a term of the form $(1 - \eta) m\, V'_{\mathrm{base}} + \eta (b - a)^2/12$, which under the truncation choice of Corollary 8.1 is itself proportional to the baseline variance. These constructions reflect the affine structure of the theoretical bounds in Lemma 8 and Corollary 8.1, and demonstrate that contamination preserves the location of the interpolation peak while amplifying predictive dispersion near $N = n$.

The likelihood contamination parameter η introduced in Lemma 1 provides an abstract model of distributional misspecification, allowing a fraction of the data-generating process to deviate arbitrarily from the assumed likelihood. To connect this formal notion with concrete data corruption, we consider an explicit misspecification experiment based on label outliers.

Experimental setup. We generate training data according to a random-feature teacher model, $y = f^*(x) + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, and introduce label misspecification via a Huber-type contamination model,

$$y = \begin{cases} f^*(x) + \varepsilon, & \text{with probability } 1 - \eta, \\ f^*(x) + \varepsilon + A, & \text{with probability } \eta, \end{cases} \tag{8}$$

where $A \gg \sigma$ is a fixed outlier amplitude. The parameter η controls the fraction of corrupted labels.

For each contamination level η, we train a ridgeless random-feature regression model, i.e. we set $\lambda = 0$, and evaluate the test mean squared error (MSE) as a function of the feature-to-dimension ratio $\psi_1 = N/d$, keeping the sample ratio $\psi_2 = n/d$ fixed.

Results. Figure 1 reports the test MSE curves under increasing levels of label contamination. In the absence of misspecification ($\eta = 0$), the estimator exhibits the familiar double descent behavior, with a sharp error peak near the interpolation threshold $N = n$. As the contamination level η increases, the height of the peak grows substantially, indicating strong noise amplification in the vicinity of interpolation.

Importantly, the location of the peak remains stable across contamination levels, while only its magnitude changes. Although Lemma 8 and Corollary 8.1 concern predictive variance rather than test MSE directly, the present behavior is qualitatively consistent with those results, together with the bias-variance decomposition above, insofar as contamination primarily amplifies the variance-dominated interpolation peak without shifting its location in ψ_1. The numerical results therefore support the interpretation of η as a robustness parameter that controls the magnitude of uncertainty under misspecification, while preserving the structural double descent phenomenon. This experiment illustrates that the imprecision-based variance envelopes derived in Section 4 are not merely abstract worst-case bounds, but are qualitatively informative about the behavior of test error under concrete data corruption when the bias contribution remains comparatively stable.

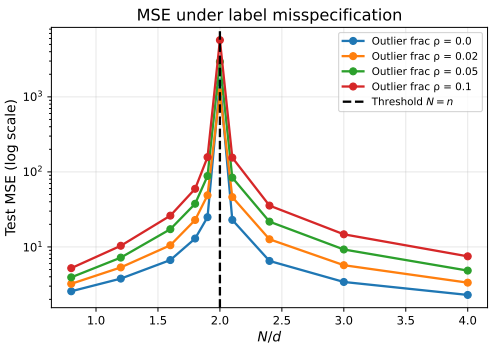


Figure 1: Test MSE as a function of N/d under increasing levels of label misspecification η. Larger contamination amplifies the interpolation peak without shifting its location, consistent with Lemma 8 and Corollary 8.1.

7 CONCLUSION

We introduced a robust Bayesian formulation of random feature regression based on Huber-style contamination sets for both prior and likelihood. By replacing a single probabilistic specification with credal sets and performing pessimistic generalized Bayesian updating, we obtained lower and upper posterior predictive that explicitly quantify worst-case predictive behavior under bounded model uncertainty.

Our main theoretical contribution is the derivation of simple and tractable bounds on predictive densities and predictive variance. These bounds act as affine transformations of the classical Bayesian RF predictive distribution, yielding uncertainty envelopes that preserve the leading-order proportional-growth asymptotics of random feature models. In particular, the interpolation-driven double-descent structure of predictive variance is retained, while its magnitude is adjusted to reflect prior and likelihood contamination. On the practical side, we introduced the Imprecise Highest Density Region (IHDR) and showed that it admits an efficient approximation via adjusted Gaussian credible intervals. This preserves the computational tractability of Bayesian random features while providing robust guarantees.

Overall, our results establish a robustness theory for Bayesian random features: predictive uncertainty remains analytically tractable and asymptotically well-behaved, yet is improved by explicit worst-case guarantees under bounded misspecification. Future work may investigate data-adaptive choices of contamination levels, extensions to kernel limits and deep feature models, and implications of robust predictive uncertainty in high-dimensional regimes.

Acknowledgements

Michele Caprio and Katerina Papagiannouli gratefully acknowledge the hospitality of the Max Planck Institute for Mathematics in the Sciences where the ideas for the present paper first took shape, and in particular of Katharina Matschke and Bernd Sturmfels. This work is dedicated to the memory of Sayan Mukherjee, whose untimely passing occurred before the paper was finalized; he is deeply missed. ChatGPT 5.5 Thinking was used to assist with proofreading.

References

Youngsoo Baek, Samuel Berchuck, and Sayan Mukherjee. Asymptotics of Bayesian Uncertainty Estimation in Random Features Regression. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=xgzkuTGBTx.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine learning practice and the classical bias–variance trade-off. *PNAS*, 2019.

James O. Berger. The robust Bayesian viewpoint. In Joseph B. Kadane, editor, *Robustness of Bayesian Analyses*. Amsterdam : North-Holland, 1984.

Michele Caprio. Optimal transport for ϵ-contaminated credal sets. In Sébastien Destercke, Alexander Erreygers, Max Nendel, Frank Riedel, and Matthias C. M. Troffaes, editors, *Proceedings of the Fourteenth International Symposium on Imprecise Probabilities: Theories and Applications*, volume 290 of *Proceedings of Machine Learning Research*, pages 33–46. PMLR, 15–18 Jul 2025. URL https://proceedings.mlr.press/v290/caprio25a.html.

Michele Caprio and Teddy Seidenfeld. Constriction for sets of probabilities. In Enrique Miranda, Ignacio Montes, Erik Quaeghebeur, and Barbara Vantaggi, editors, *Proceedings of the Thirteenth International Symposium on Imprecise Probability: Theories and Applications*, volume 215 of *Proceedings of Machine Learning Research*, pages 84–95. PMLR, 11–14 Jul 2023. URL https://proceedings.mlr.press/v215/caprio23b.html.

Michele Caprio, Souradeep Dutta, Kuk Jin Jang, Vivian Lin, Radoslav Ivanov, Oleg Sokolsky, and Insup Lee. Credal bayesian deep learning. *Transactions on Machine Learning Research*, 2024a. ISSN 2835-8856. URL https://openreview.net/forum?id=4NHF9AC5ui.

Michele Caprio, Yusuf Sale, Eyke Hüllermeier, and Insup Lee. A Novel Bayes' Theorem for Upper Probabilities. In Fabio Cuzzolin and Maryam Sultana, editors, *Epistemic Uncertainty in Artificial Intelligence*, pages 1–12, Cham, 2024b. Springer Nature Switzerland.

Simone Cerreia-Vioglio, Fabio Maccheroni, and Massimo Marinacci. Ergodic theorems for lower probabilities. *Proceedings of the American Mathematical Society*, 144: 3381–3396, 2015.

Mengqi Chen, Thomas B. Berrett, Theodoros Damoulas, and Michele Caprio. Bulk-calibrated credal ambiguity sets: Fast, tractable decision making under out-of-sample contamination, 2026. URL https://arxiv.org/abs/2601.21324.

Frank P. A. Coolen. Imprecise highest density regions related to intervals of measures. *Memorandum COSOR*, 9254, 1992.

Charita Dellaporta, Patrick O'Hara, and Theodoros Damoulas. Distributionally robust optimisation with bayesian ambiguity sets. In *NeurIPS 2024 Workshop on Bayesian Decision-making and Uncertainty*, 2024. URL https://openreview.net/forum?id=woNEqCWJ9g.

Ruobin Gong and Xiao-Li Meng. Judicious judgment meets unsettling updating: dilation, sure loss, and Simpson's paradox. *Statistical Science*, 36(2):169–190, 2021.

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

Peter J. Huber and Elvezio M. Ronchetti. *Robust statistics*. Wiley Series in Probability and Statistics. Hoboken, New Jersey : Wiley, 2nd edition, 2009.

Rob J. Hyndman. Computing and graphing highest density regions. *The American Statistician*, 50(2):120–126, 1996.

Isaac Levi. *The Enterprise of Knowledge*. London, UK : MIT Press, 1980.

Francisco Louzada, Paulo H. Ferreira, and Diego C. Nascimento. *Spike-and-Slab Priors and Their Applications*, pages 1–8. John Wiley & Sons, Ltd, 2023. ISBN 9781118445112. doi: https://doi.org/10.1002/9781118445112.stat08417. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118445112.stat08417.

Massimo Marinacci and Luigi Montrucchio. Introduction to the mathematics of ambiguity. In Itzhak Gilboa, editor, *Uncertainty in economic theory: a collection of essays in honor of David Schmeidler's 65th birthday*. London : Routledge, 2004.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. *Communications on Pure and Applied Mathematics*, 2022.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007. URL https://proceedings.neurips.cc/paper_files/paper/2007/file/013a006f03dbc5392effeb8f18fda755-Paper.pdf.

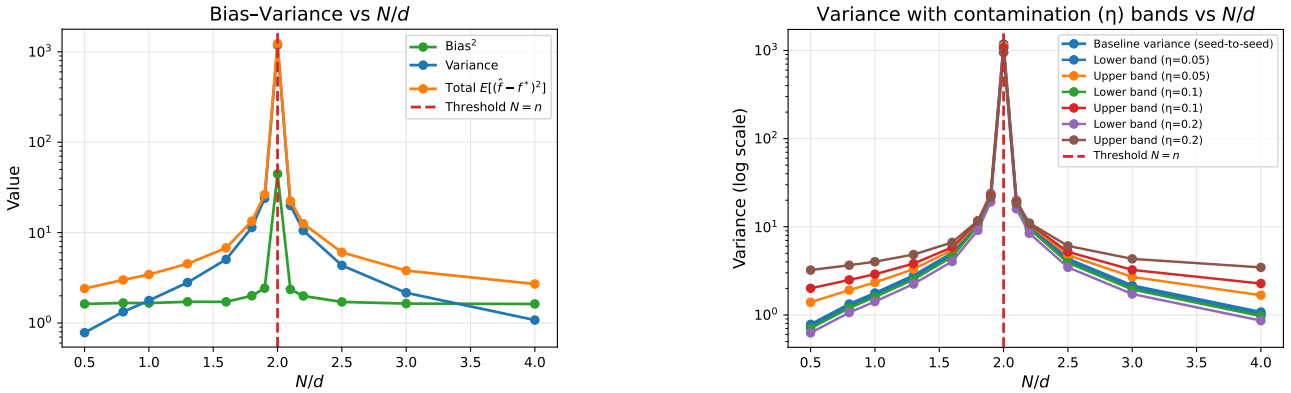
Patrick Suppes and Mario Zanotti. On using random relations to generate upper and lower probabilities. *Synthese*, 36(4):427–440, 1977. ISSN 00397857, 15730964. URL http://www.jstor.org/stable/20115240.

Matthias C.M. Troffaes and Gert de Cooman. *Lower Previsions*. Chichester, United Kingdom: John Wiley and Sons, 2014.

Peter Walley. *Statistical Reasoning with Imprecise Probabilities*, volume 42 of *Monographs on Statistics and Applied Probability*. London : Chapman and Hall, 1991.

Larry A. Wasserman and Joseph B. Kadane. Bayes' theorem for Choquet capacities. *The Annals of Statistics*, 18(3): 1328–1339, 1990.

A BIAS-VARIANCE AND CONTAMINATION PLOTS



(a) Bias-variance decomposition

(b) Variance with η-contamination bands

Figure 2: Bias-variance mechanism and robustness to likelihood contamination. *Left:* Bias^2, variance, and total error as functions of $\psi_1 = N/d$, showing that the double descent peak is variance-driven. *Right:* Contamination-induced variance envelopes for $\eta \in \{0.05, 0.1, 0.2\}$, illustrating the affine bounds of Lemma 8 and the persistence of the interpolation peak predicted by Corollary 8.1.

B PROOFS OF OUR RESULTS

Proof of Lemma 2. Let us begin with the lower density. Fix any $A \in \mathcal{B}(\mathbb{R}^N) \setminus \{\mathbb{R}^N\}$. By Wasserman and Kadane [1990, Example 3], we know that $\underline{P}(A) = (1-\epsilon)P^{\mathcal{N}}(A) = (1-\epsilon)P^{\mathcal{N}}(A) + \epsilon Q_{\star A}(A)$. In turn, we have that

$$\begin{aligned}\underline{P}(A) &= (1-\epsilon)\int_A p^{\mathcal{N}}(\boldsymbol{a})\mathrm{d}\boldsymbol{a} + \epsilon\int_A q_{\star A}(\boldsymbol{a})\mathrm{d}\boldsymbol{a} \\ &= \int_A \left[(1-\epsilon)p^{\mathcal{N}}(\boldsymbol{a}) + \epsilon q_{\star A}(\boldsymbol{a})\right]\mathrm{d}\boldsymbol{a} = \int_A \underline{p}(\boldsymbol{a})\mathrm{d}\boldsymbol{a}.\end{aligned}$$

For the upper density, fix any $A \in \mathcal{B}(\mathbb{R}^N) \setminus \{\emptyset\}$. By Wasserman and Kadane [1990, Example 3], we know that $\overline{P}(A) = (1-\epsilon)P^{\mathcal{N}}(A) + \epsilon = (1-\epsilon)P^{\mathcal{N}}(A) + \epsilon Q^{\star}_A(A)$. In turn, we have that

$$\begin{aligned}\overline{P}(A) &= (1-\epsilon)\int_A p^{\mathcal{N}}(\boldsymbol{a})\mathrm{d}\boldsymbol{a} + \epsilon\int_A q^{\star}_A(\boldsymbol{a})\mathrm{d}\boldsymbol{a} \\ &= \int_A \left[(1-\epsilon)p^{\mathcal{N}}(\boldsymbol{a}) + \epsilon q^{\star}_A(\boldsymbol{a})\right]\mathrm{d}\boldsymbol{a} = \int_A \overline{p}(\boldsymbol{a})\mathrm{d}\boldsymbol{a}.\end{aligned}$$

□

Proof of Theorem 3. In what follows, we take $\underline{p}_{\boldsymbol{a}}$ to denote the A-independent lower envelope subdensity $(1-\epsilon)p^{\mathcal{N}}(\boldsymbol{a})$. Throughout this proof, consistently with the absolutely continuous restriction discussed after Lemma 2, we restrict the likelihood contaminations to a subclass of bounded spike densities centered at the observed data $\mathbf{y}$, and we take this spike to be fixed, i.e. independent of the measurable set under consideration, of the level t in the Choquet representation, and of $\boldsymbol{a}$. Hence, in the likelihood calculations below, we write $\overline{\ell}_{\boldsymbol{a}}(\boldsymbol{y}) = (1-\eta)\ell^{\mathcal{N}}_{\boldsymbol{a}}(\boldsymbol{y}) + \eta\delta_{\mathbf{y}}(\boldsymbol{y})$, where $\delta_{\mathbf{y}}$ is a bounded spike density centered at $\mathbf{y}$. Let $\underline{\ell}_{\boldsymbol{a}}(\boldsymbol{y}) \equiv \underline{p}(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\Theta}, \boldsymbol{a})$, and similarly for $\overline{\ell}_{\boldsymbol{a}}(\boldsymbol{y})$. Under the absolutely continuous restriction discussed after Lemma 2, we work with the following density-level representation of the pessimistic generalized Bayes update (6),

$$\underline{p}(\boldsymbol{a} \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\Theta}) = \frac{\underline{p}(\boldsymbol{a})\underline{\ell}_{\boldsymbol{a}}(\boldsymbol{y})}{\sup_{P\in\mathcal{P}_\epsilon}\int \overline{\ell}_{\boldsymbol{a}}(\boldsymbol{y})P(d\boldsymbol{a})}. \tag{9}$$

We also have that $\overline{P}$ is 2-alternating, i.e. $\overline{P}(A\cup B) \leq \overline{P}(A) + \overline{P}(B) - \overline{P}(A\cap B)$, for all $A, B \in \mathcal{B}(\mathbb{R}^N)$ [Wasserman and Kadane, 1990, Caprio et al., 2024b]. This means that the upper expectation $\sup_{P\in\mathcal{P}_\epsilon}\int \overline{\ell}_{\boldsymbol{a}}(\boldsymbol{y})P(d\boldsymbol{a})$ coincides with

the Choquet integral $\int \overline{\ell}_{\boldsymbol{a}}(\boldsymbol{y})\overline{P}(d\boldsymbol{a})$ [Cerreia-Vioglio et al., 2015], making computations easier. Let us focus on such an integration. We have that

$$\int \overline{\ell}_{\boldsymbol{a}}(\boldsymbol{y})\overline{P}(d\boldsymbol{a}) = \int_0^{\infty} \overline{P}\left(\left\{\boldsymbol{a} : \overline{\ell}_{\boldsymbol{a}}(\boldsymbol{y}) \geq t\right\}\right) \mathrm{d}t \tag{10}$$

$$= \int_0^{\text{ somewhere}} \overline{P}\left(\left\{\boldsymbol{a} : \overline{\ell}_{\boldsymbol{a}}(\boldsymbol{y}) \geq t\right\}\right) \mathrm{d}t \tag{11}$$

$$= \int_0^{\text{ somewhere}} \left[(1-\epsilon)P^{\mathcal{N}}\left(\left\{\boldsymbol{a} : \overline{\ell}_{\boldsymbol{a}}(\boldsymbol{y}) \geq t\right\}\right) + \epsilon Q^{\star}_{\left\{\boldsymbol{a}:\overline{\ell}_{\boldsymbol{a}}(\boldsymbol{y})\geq t\right\}}\left(\left\{\boldsymbol{a} : \overline{\ell}_{\boldsymbol{a}}(\boldsymbol{y}) \geq t\right\}\right)\right] \mathrm{d}t \tag{12}$$

$$= (1-\epsilon)\int_0^{\text{ somewhere}} P^{\mathcal{N}}\left(\left\{\boldsymbol{a} : \overline{\ell}_{\boldsymbol{a}}(\boldsymbol{y}) \geq t\right\}\right) \mathrm{d}t + \epsilon \int_0^{\text{ somewhere}} \underbrace{Q^{\star}_{\left\{\boldsymbol{a}:\overline{\ell}_{\boldsymbol{a}}(\boldsymbol{y})\geq t\right\}}\left(\left\{\boldsymbol{a} : \overline{\ell}_{\boldsymbol{a}}(\boldsymbol{y}) \geq t\right\}\right)}_{=1 \text{ for all } t} \mathrm{d}t \tag{13}$$

$$= (1-\epsilon)\int \overline{\ell}_{\boldsymbol{a}}(\boldsymbol{y}) p^{\mathcal{N}}(\boldsymbol{a})\mathrm{d}\boldsymbol{a} + \underbrace{\epsilon \int_0^{\text{ somewhere}} \mathrm{d}t}_{=\epsilon\cdot \text{somewhere} =: c}$$

$$= (1-\epsilon)\int \left[(1-\eta)\ell^{\mathcal{N}}_{\boldsymbol{a}}(\boldsymbol{y}) + \eta\delta_{\boldsymbol{y}}\right] p^{\mathcal{N}}(\boldsymbol{a})\mathrm{d}\boldsymbol{a} + c \tag{14}$$

$$= (1-\epsilon)(1-\eta)\int \ell^{\mathcal{N}}_{\boldsymbol{a}}(\boldsymbol{y}) p^{\mathcal{N}}(\boldsymbol{a})\mathrm{d}\boldsymbol{a} + (1-\epsilon)\eta\delta_{\boldsymbol{y}}\int p^{\mathcal{N}}(\boldsymbol{a})\mathrm{d}\boldsymbol{a} + c$$

$$= (1-\epsilon)(1-\eta)p(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\Theta}) + (1-\epsilon)\eta\delta_{\boldsymbol{y}} + c, \tag{15}$$

where (10) comes from Troffaes and de Cooman [2014, Proposition C.3] and Marinacci and Montrucchio [2004, Equation (11)]; (11) comes from $\overline{\ell}_{\boldsymbol{a}}(\boldsymbol{y})$ being bounded and having set $\text{somewhere} := \sup_{\boldsymbol{a}} \overline{\ell}_{\boldsymbol{a}}(\boldsymbol{y})$; (12) comes from Lemma 2; (13) comes from the additivity of the integral operator and Lemma 2; in (14), the spike density term $\delta_{\boldsymbol{y}}$ is at the observed data $\mathbf{y}$; (15) comes from $p(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\Theta}) = \int \ell^{\mathcal{N}}_{\boldsymbol{a}}(\boldsymbol{y}) p^{\mathcal{N}}(\boldsymbol{a})\mathrm{d}\boldsymbol{a}$ being the marginal likelihood in the classical Bayesian RF model, and $\delta_{\boldsymbol{y}} \int p^{\mathcal{N}}(\boldsymbol{a})d\boldsymbol{a} = \delta_{\boldsymbol{y}}$ because $\delta_{\boldsymbol{y}}$ does not depend on $\boldsymbol{a}$, and $\int p^{\mathcal{N}}(\boldsymbol{a})d\boldsymbol{a} = 1$ because $p^{\mathcal{N}}(\boldsymbol{a})$ is a proper density.

In turn, by Lemma 2 and by substituting (15) in (9), we obtain

$$\underline{p}(\boldsymbol{a} \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\Theta}) = \frac{(1-\epsilon)p^{\mathcal{N}}(\boldsymbol{a})(1-\eta)\ell^{\mathcal{N}}_{\boldsymbol{a}}(\boldsymbol{y})}{(1-\epsilon)(1-\eta)\int \ell^{\mathcal{N}}_{\boldsymbol{a}}(\boldsymbol{y}) p^{\mathcal{N}}(\boldsymbol{a})\mathrm{d}\boldsymbol{a} + (1-\epsilon)\eta\delta_{\boldsymbol{y}} + c}.$$

For proper measurable sets $A \in \mathcal{B}(\mathbb{R}^N) \setminus \{\mathbb{R}^N\}$, this expression is understood as the subdensity representation associated with the updated lower probability, while for $A = \mathbb{R}^N$ one must appeal to the proper setwise characterization recalled in Lemma 1. In particular, in the derivation of the predictive bound below, we use this subdensity representation as an auxiliary density-level device for bounding the lower posterior predictive envelope, rather than as a globally normalized conditional density on all of $\mathbb{R}^N$. It can be routinely checked that, on proper measurable sets and together with the full-space convention from Lemma 1, the corresponding updated set function defines a well-behaved lower probability. That is, for all $B \in \mathcal{B}(\mathbb{R}^n)$, it satisfies

(i) $\underline{P}(\boldsymbol{a} \in \emptyset \mid \boldsymbol{X}, \boldsymbol{\Theta}, \boldsymbol{y} \in B) = 0$,

(ii) $\underline{P}(\boldsymbol{a} \in \mathbb{R}^N \mid \boldsymbol{X}, \boldsymbol{\Theta}, \boldsymbol{y} \in B) = 1$,

(iii) $A_1 \subseteq A_2 \implies \underline{P}(\boldsymbol{a} \in A_1 \mid \boldsymbol{X}, \boldsymbol{\Theta}, \boldsymbol{y} \in B) \leq \underline{P}(\boldsymbol{a} \in A_2 \mid \boldsymbol{X}, \boldsymbol{\Theta}, \boldsymbol{y} \in B)$,

(iv) $\underline{P}(\boldsymbol{a} \in A_1 \sqcup A_2 \mid \boldsymbol{X}, \boldsymbol{\Theta}, \boldsymbol{y} \in B) \geq \underline{P}(\boldsymbol{a} \in A_1 \mid \boldsymbol{X}, \boldsymbol{\Theta}, \boldsymbol{y} \in B) + \underline{P}(\boldsymbol{a} \in A_2 \mid \boldsymbol{X}, \boldsymbol{\Theta}, \boldsymbol{y} \in B)$.

Suppose now we are given a new input $\tilde{\boldsymbol{x}}$. We then focus on the lower posterior predictive density $\underline{\ell}_{\text{pred}}(\tilde{y}) \equiv \underline{p}(\tilde{y} \mid \tilde{\boldsymbol{x}}, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\Theta})$. Let us denote the likelihood $\underline{\ell}_{\boldsymbol{a}}(\tilde{y}) \equiv \underline{p}(\tilde{y} \mid \tilde{\boldsymbol{x}}, \boldsymbol{\Theta}, \boldsymbol{a})$ for ease of notation. Using a reasoning similar to that above, we write

$$
\begin{aligned}
\underline{\ell}_{\text{pred}}(\tilde{y}) &= \inf_{P \in \mathcal{M}(\underline{P}(\cdot \mid \boldsymbol{X}, \boldsymbol{\Theta}, \boldsymbol{y} \in B))} \int \underline{\ell}_{\boldsymbol{a}}(\tilde{y}) P(\mathrm{d}\boldsymbol{a} \mid \boldsymbol{X}, \boldsymbol{\Theta}, \boldsymbol{y} \in B) \\
&= \int \underline{\ell}_{\boldsymbol{a}}(\tilde{y}) \underline{P}(\mathrm{d}\boldsymbol{a} \mid \boldsymbol{X}, \boldsymbol{\Theta}, \boldsymbol{y} \in B) && (16) \\
&= \int \underline{\ell}_{\boldsymbol{a}}(\tilde{y}) \underline{p}(\boldsymbol{a} \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\Theta}) \mathrm{d}\boldsymbol{a} && (17) \\
&= \int (1-\eta) \ell^{\mathcal{N}}_{\boldsymbol{a}}(\tilde{y}) \left[\frac{(1-\epsilon) p^{\mathcal{N}}(\boldsymbol{a})(1-\eta) \ell^{\mathcal{N}}_{\boldsymbol{a}}(\boldsymbol{y})}{(1-\epsilon)(1-\eta) \int \ell^{\mathcal{N}}_{\boldsymbol{a}}(\boldsymbol{y}) p^{\mathcal{N}}(\boldsymbol{a}) \mathrm{d}\boldsymbol{a} + (1-\epsilon)\eta \delta_{\boldsymbol{y}} + c} \right] \mathrm{d}\boldsymbol{a} \\
&= \int \frac{(1-\epsilon)(1-\eta)^2 p^{\mathcal{N}}(\boldsymbol{a}) \ell^{\mathcal{N}}_{\boldsymbol{a}}(\boldsymbol{y}) \ell^{\mathcal{N}}_{\boldsymbol{a}}(\tilde{y})}{(1-\epsilon)(1-\eta) \int \ell^{\mathcal{N}}_{\boldsymbol{a}}(\boldsymbol{y}) p^{\mathcal{N}}(\boldsymbol{a}) \mathrm{d}\boldsymbol{a} + (1-\epsilon)\eta \delta_{\boldsymbol{y}} + c} \mathrm{d}\boldsymbol{a} \\
&= \frac{(1-\epsilon)(1-\eta)^2}{(1-\epsilon)(1-\eta) \int \ell^{\mathcal{N}}_{\boldsymbol{a}}(\boldsymbol{y}) p^{\mathcal{N}}(\boldsymbol{a}) \mathrm{d}\boldsymbol{a} + (1-\epsilon)\eta \delta_{\boldsymbol{y}} + c} \int p^{\mathcal{N}}(\boldsymbol{a}) \ell^{\mathcal{N}}_{\boldsymbol{a}}(\boldsymbol{y}) \ell^{\mathcal{N}}_{\boldsymbol{a}}(\tilde{y}) \mathrm{d}\boldsymbol{a} \\
&= \frac{(1-\epsilon)(1-\eta)^2 \int \ell^{\mathcal{N}}_{\boldsymbol{a}}(\boldsymbol{y}) p^{\mathcal{N}}(\boldsymbol{a}) \mathrm{d}\boldsymbol{a}}{(1-\epsilon)(1-\eta) \int \ell^{\mathcal{N}}_{\boldsymbol{a}}(\boldsymbol{y}) p^{\mathcal{N}}(\boldsymbol{a}) \mathrm{d}\boldsymbol{a} + (1-\epsilon)\eta \delta_{\boldsymbol{y}} + c} \int \ell^{\mathcal{N}}_{\boldsymbol{a}}(\tilde{y}) \underbrace{\frac{p^{\mathcal{N}}(\boldsymbol{a}) \ell^{\mathcal{N}}_{\boldsymbol{a}}(\boldsymbol{y})}{\int p^{\mathcal{N}}(\boldsymbol{a}) \ell^{\mathcal{N}}_{\boldsymbol{a}}(\boldsymbol{y}) \mathrm{d}\boldsymbol{a}}}_{\text{this ratio is } = p^{\mathcal{N}}(\boldsymbol{a} \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\Theta})} \mathrm{d}\boldsymbol{a} && (18) \\
&= \frac{(1-\epsilon)(1-\eta)^2 \int \ell^{\mathcal{N}}_{\boldsymbol{a}}(\boldsymbol{y}) p^{\mathcal{N}}(\boldsymbol{a}) \mathrm{d}\boldsymbol{a}}{(1-\epsilon)(1-\eta) \int \ell^{\mathcal{N}}_{\boldsymbol{a}}(\boldsymbol{y}) p^{\mathcal{N}}(\boldsymbol{a}) \mathrm{d}\boldsymbol{a} + (1-\epsilon)\eta \delta_{\boldsymbol{y}} + c} \underbrace{p(\tilde{y} \mid \tilde{\boldsymbol{x}}, \boldsymbol{y}, \boldsymbol{X}, \Theta)}_{\equiv \ell_{\text{pred}}(\tilde{y})} && (19) \\
&\leq (1-\eta) \ell_{\text{pred}}(\tilde{y}), && (20)
\end{aligned}
$$

where (16) comes from the posterior lower probability $\underline{P}(\cdot \mid \boldsymbol{X}, \boldsymbol{\Theta}, \boldsymbol{y} \in B)$ being 2-monotone [Wasserman and Kadane, 1990, Caprio et al., 2024b]; (17) comes from writing the Choquet integral in terms of the superlevel sets of $\underline{\ell}_{\boldsymbol{a}}(\tilde{y})$, namely the sets $\{\boldsymbol{a} \in \mathbb{R}^N : \underline{\ell}_{\boldsymbol{a}}(\tilde{y}) \geq t\}$, $t > 0$, and observing that these are proper measurable subsets of $\mathbb{R}^N$; in (18), we denoted by $p^{\mathcal{N}}(\boldsymbol{a} \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\Theta})$ the (classical) posterior derived from (3) and (4); in (19), we denoted by $p(\tilde{y} \mid \tilde{\boldsymbol{x}}, \boldsymbol{y}, \boldsymbol{X}, \Theta)$ the (classical) posterior predictive distribution, derived from (3) and (4), and the new input $\tilde{\boldsymbol{x}}$.

Equation (20) gives an upper bound for $\underline{\ell}_{\text{pred}}(\tilde{y})$. Moreover, if $(1-\epsilon)\eta\delta_{\boldsymbol{y}} + c$ is negligible compared to $(1-\epsilon)(1-\eta) \int \ell^{\mathcal{N}}_{\boldsymbol{a}}(\boldsymbol{y}) p^{\mathcal{N}}(\boldsymbol{a}) d\boldsymbol{a}$, then $\underline{\ell}_{\text{pred}}(\tilde{y}) \approx (1-\eta)\ell_{\text{pred}}(\tilde{y})$. □

Proof of Corollary 3.1. Similarly to the proof of Theorem 3, under the absolutely continuous bounded-spike subclass discussed after Lemma 2, and restricting attention to contaminating prior densities for which a global upper density representation exists, we work with the following density-level representation of the optimistic generalized Bayes update,

$$
\overline{p}(\boldsymbol{a} \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\Theta}) = \frac{\overline{p}(\boldsymbol{a}) \overline{\ell}_{\boldsymbol{a}}(\boldsymbol{y})}{\inf_{P \in \mathcal{P}_\epsilon} \int \underline{\ell}_{\boldsymbol{a}}(\boldsymbol{y}) P(d\boldsymbol{a})}, \tag{21}
$$

where $\overline{p}(\boldsymbol{a})$ denotes the A-independent upper density, when it exists.

Fix any probability density u_n on $\mathbb{R}^n$ and the density u on $\mathbb{R}$ appearing in the statement of Corollary 3.1. Throughout this proof, interpret the likelihood contamination term in $\overline{\ell}_{\boldsymbol{a}}(\boldsymbol{y})$ as $\eta u_n(\boldsymbol{y})$, and the one in $\overline{\ell}_{\boldsymbol{a}}(\tilde{y})$ as $\eta u(\tilde{y})$.

Let us focus on the integral at the denominator. We know that $\underline{P} \coloneqq \inf_{P \in \mathcal{P}_\epsilon} P$ is 2-monotone, i.e. $\underline{P}(A \cup B) \geq \underline{P}(A) + \underline{P}(B) - \underline{P}(A \cap B)$, for all $A, B \in \mathcal{B}(\mathbb{R}^N)$ – see e.g. Wasserman and Kadane [1990], Caprio et al. [2024b]. This means that the lower expectation $\inf_{P \in \mathcal{P}_\epsilon} \int \underline{\ell}_{\boldsymbol{a}}(\boldsymbol{y}) P(d\boldsymbol{a})$ coincides with the Choquet integral $\int \underline{\ell}_{\boldsymbol{a}}(\boldsymbol{y}) \underline{P}(d\boldsymbol{a})$ [Cerreia-Vioglio et al.,

2015]. In turn, we have that

$$\begin{aligned}
\int \underline{\ell}_{\boldsymbol{a}}(\boldsymbol{y})\underline{P}(d\boldsymbol{a}) &= \int_0^{\infty} \underline{P}\left(\{\boldsymbol{a} : \underline{\ell}_{\boldsymbol{a}}(\boldsymbol{y}) \geq t\}\right) \mathrm{d}t \\
&= \int_0^{\overline{\ell}'} \left[(1-\epsilon)P^{\mathcal{N}}\left(\{\boldsymbol{a} : \underline{\ell}_{\boldsymbol{a}}(\boldsymbol{y}) \geq t\}\right) + \epsilon Q_{\star\left\{\boldsymbol{a}:\underline{\ell}_{\boldsymbol{a}}(\boldsymbol{y})\geq t\right\}}\left(\{\boldsymbol{a} : \underline{\ell}_{\boldsymbol{a}}(\boldsymbol{y}) \geq t\}\right)\right] \mathrm{d}t \\
&= (1-\epsilon)\int_0^{\overline{\ell}'} P^{\mathcal{N}}\left(\{\boldsymbol{a} : \underline{\ell}_{\boldsymbol{a}}(\boldsymbol{y}) \geq t\}\right) \mathrm{d}t + \epsilon \int_0^{\overline{\ell}'} \underbrace{Q_{\star\left\{\boldsymbol{a}:\underline{\ell}_{\boldsymbol{a}}(\boldsymbol{y})\geq t\right\}}\left(\{\boldsymbol{a} : \underline{\ell}_{\boldsymbol{a}}(\boldsymbol{y}) \geq t\}\right)}_{=0 \text{ for all } t} \mathrm{d}t \\
&= (1-\epsilon)\int \underline{\ell}_{\boldsymbol{a}}(\boldsymbol{y}) p^{\mathcal{N}}(\boldsymbol{a})\mathrm{d}\boldsymbol{a} \\
&= (1-\epsilon)\int (1-\eta)\ell_{\boldsymbol{a}}^{\mathcal{N}}(\boldsymbol{y}) p^{\mathcal{N}}(\boldsymbol{a})\mathrm{d}\boldsymbol{a} \\
&= (1-\eta)(1-\epsilon)\underbrace{\int \ell_{\boldsymbol{a}}^{\mathcal{N}}(\boldsymbol{y}) p^{\mathcal{N}}(\boldsymbol{a})\mathrm{d}\boldsymbol{a}}_{=p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\Theta})},
\end{aligned}$$

where $\overline{\ell}' \coloneqq \sup_{\boldsymbol{a}} \underline{\ell}_{\boldsymbol{a}}(\boldsymbol{y})$. So, we can rewrite (21) as

$$\begin{aligned}
\overline{p}(\boldsymbol{a} \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\Theta}) &= \frac{\overline{p}(\boldsymbol{a})\overline{\ell}_{\boldsymbol{a}}(\boldsymbol{y})}{(1-\eta)(1-\epsilon)\int \ell_{\boldsymbol{a}}^{\mathcal{N}}(\boldsymbol{y}) p^{\mathcal{N}}(\boldsymbol{a})\mathrm{d}\boldsymbol{a}} \\
&\geq \frac{(1-\epsilon)p^{\mathcal{N}}(\boldsymbol{a})\left[(1-\eta)\ell_{\boldsymbol{a}}^{\mathcal{N}}(\boldsymbol{y}) + \eta u_n(\boldsymbol{y})\right]}{(1-\eta)(1-\epsilon)\int \ell_{\boldsymbol{a}}^{\mathcal{N}}(\boldsymbol{y}) p^{\mathcal{N}}(\boldsymbol{a})\mathrm{d}\boldsymbol{a}} \\
&\geq \frac{(1-\eta)(1-\epsilon)p^{\mathcal{N}}(\boldsymbol{a})\ell_{\boldsymbol{a}}^{\mathcal{N}}(\boldsymbol{y})}{(1-\eta)(1-\epsilon)\int \ell_{\boldsymbol{a}}^{\mathcal{N}}(\boldsymbol{y}) p^{\mathcal{N}}(\boldsymbol{a})\mathrm{d}\boldsymbol{a}} = p(\boldsymbol{a} \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\Theta}).
\end{aligned}$$

Since optimistic generalized Bayes preserves concavity of upper probabilities, the updated upper probability is 2-alternating [Wasserman and Kadane, 1990, Caprio et al., 2024b]. Hence, interpreting the upper posterior predictive prevision as the natural extension of the updated upper probability, it coincides with the Choquet integral with respect to that updated upper probability. For proper measurable sets $A \in \mathcal{B}(\mathbb{R}^N) \setminus \{\emptyset\}$, the expression above is understood as the superdensity representation associated with the updated upper probability, while for $A = \emptyset$ one appeals to the proper setwise characterization in Lemma 1. In particular, since for every $t > 0$ the level sets $\{\boldsymbol{a} \in \mathbb{R}^N : \overline{\ell}_{\boldsymbol{a}}(\tilde{y}) \geq t\}$ are proper measurable subsets of $\mathbb{R}^N$, this superdensity representation is sufficient for the Choquet calculation below; hence, with a slight abuse of notation, we may write the corresponding Choquet integral in superdensity form. We can then turn our attention to the upper posterior predictive density. We have that

$$\begin{aligned}
\overline{\ell}_{\text{pred}}(\tilde{y}) &= \int \overline{\ell}_{\boldsymbol{a}}(\tilde{y})\overline{p}(\boldsymbol{a} \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\Theta})\mathrm{d}\boldsymbol{a} \\
&\geq \int \overline{\ell}_{\boldsymbol{a}}(\tilde{y}) p(\boldsymbol{a} \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\Theta})\mathrm{d}\boldsymbol{a} \\
&= \int \left[(1-\eta)\ell_{\boldsymbol{a}}^{\mathcal{N}}(\tilde{y}) + \eta u(\tilde{y})\right] p(\boldsymbol{a} \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\Theta})\mathrm{d}\boldsymbol{a} \\
&= (1-\eta)\int \ell_{\boldsymbol{a}}^{\mathcal{N}}(\tilde{y}) p(\boldsymbol{a} \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\Theta})\mathrm{d}\boldsymbol{a} + \eta u(\tilde{y}) \underbrace{\int p(\boldsymbol{a} \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\Theta})\mathrm{d}\boldsymbol{a}}_{=1} \\
&= (1-\eta)\underbrace{p^{\mathcal{N}}(\tilde{y} \mid \tilde{\boldsymbol{x}}, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\Theta})}_{\equiv \ell_{\text{pred}}(\tilde{y})} + \eta u(\tilde{y}),
\end{aligned}$$

where the first equality is the Choquet integral written in superdensity form, justified by the preceding paragraph, and the inequality follows from the pointwise bound $\overline{p}(\boldsymbol{a} \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\Theta}) \geq p(\boldsymbol{a} \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\Theta})$ together with the nonnegativity of $\overline{\ell}_{\boldsymbol{a}}(\tilde{y})$. The last equality gives a lower bound for $\overline{\ell}_{\text{pred}}(\tilde{y})$. Moreover, for ϵ, η sufficiently small, the upper posterior can be well

approximated by the classical posterior whenever the contaminating contributions are negligible. Indeed, for any choice of contaminating prior density q and contaminating likelihood density u_n, write

$$\overline{p}(\boldsymbol{a}) = (1-\epsilon)p^{\mathcal{N}}(\boldsymbol{a}) + \epsilon q(\boldsymbol{a}), \qquad \overline{\ell}_{\boldsymbol{a}}(\boldsymbol{y}) = (1-\eta)\ell^{\mathcal{N}}_{\boldsymbol{a}}(\boldsymbol{y}) + \eta u_n(\boldsymbol{y}).$$

Then,

$$\overline{p}(\boldsymbol{a})\overline{\ell}_{\boldsymbol{a}}(\boldsymbol{y}) = (1-\epsilon)(1-\eta)p^{\mathcal{N}}(\boldsymbol{a})\ell^{\mathcal{N}}_{\boldsymbol{a}}(\boldsymbol{y}) + \epsilon(1-\eta)q(\boldsymbol{a})\ell^{\mathcal{N}}_{\boldsymbol{a}}(\boldsymbol{y}) + (1-\epsilon)\eta p^{\mathcal{N}}(\boldsymbol{a})u_n(\boldsymbol{y}) + \epsilon\eta q(\boldsymbol{a})u_n(\boldsymbol{y}).$$

If the last three terms are small compared to the first one (e.g. in $L^1(d\boldsymbol{a})$ over $\boldsymbol{a}$), then the resulting upper posterior is close to the classical posterior, and consequently $\overline{\ell}_{\text{pred}}(\tilde{y}) \approx (1-\eta)\ell_{\text{pred}}(\tilde{y}) + \eta u(\tilde{y})$. □

Proof of Proposition 5. Recall the definition of a β-level (precise) Highest Density Region for the probability measure P_{pred} associated with pdf ℓ_{pred}, denoted by $R_\beta[P_{\text{pred}}]$ [Hyndman, 1996]. It is the element of $\mathcal{B}(\mathbb{R})$ that satisfies two properties, (i) $P_{\text{pred}}\left(\{\tilde{y} \in R_\beta[P_{\text{pred}}]\}\right) \geq \beta$, and (ii) $\int_{R_\beta[P_{\text{pred}}]} \mathrm{d}\tilde{y}$ is a minimum. Let

$$S \coloneqq \text{IR}_\alpha[\mathcal{M}(\underline{P}_{\text{pred}})].$$

If $\alpha = 0$, then necessarily $S = \mathbb{R}$ and $\beta = 1$, so $R_\beta[P_{\text{pred}}] = \mathbb{R}$ and the claim is immediate. Hence, in the remainder of the proof, assume $0 < \alpha \leq 1$, so that S is a proper measurable set. By definition of IHDR,

$$\underline{P}_{\text{pred}}\left(\{\tilde{y} \in S\}\right) = 1 - \alpha.$$

By Theorem 3,

$$\underline{P}_{\text{pred}}\left(\{\tilde{y} \in S\}\right) = \int_S \underline{\ell}_{\text{pred}}(\tilde{y})\mathrm{d}\tilde{y} \leq (1-\eta)\int_S \ell_{\text{pred}}(\tilde{y})\mathrm{d}\tilde{y} = (1-\eta)P_{\text{pred}}\left(\{\tilde{y} \in S\}\right). \tag{22}$$

Since $P_{\text{pred}}(S) \leq 1$, (22) implies

$$1 - \alpha = \underline{P}_{\text{pred}}(S) \leq (1-\eta)P_{\text{pred}}(S) \leq 1 - \eta,$$

hence necessarily $\alpha \geq \eta$. Therefore, under the existence assumption for the IHDR and for $\eta < 1$, we have

$$\beta = \min\left\{1, \frac{1-\alpha}{1-\eta}\right\} = \frac{1-\alpha}{1-\eta}.$$

In turn,

$$P_{\text{pred}}\left(\{\tilde{y} \in S\}\right) \geq \frac{1-\alpha}{1-\eta} =: \beta.$$

Thus S is feasible for the optimization problem defining $R_\beta[P_{\text{pred}}]$. By minimality of $R_\beta[P_{\text{pred}}]$,

$$\int_{R_\beta[P_{\text{pred}}]} \mathrm{d}\tilde{y} \leq \int_S \mathrm{d}\tilde{y}.$$

Since ℓ_{pred} is a Normal distribution, it is symmetric and unimodal, and hence $R_\beta[P_{\text{pred}}]$ coincides with the classical notion of credible interval [Hyndman, 1996, Caprio et al., 2024a]. □

Proof of Proposition 6. We have that

$$\begin{aligned}
\mathbb{V}_{\underline{\ell}_{\text{pred}}}(\tilde{Y}) &\coloneqq \min_{\zeta\in\mathbb{R}} \mathbb{E}_{\underline{\ell}_{\text{pred}}}[(\tilde{Y}-\zeta)^2] \\
&= \min_{\zeta\in\mathbb{R}} \int_{\mathbb{R}} (\tilde{y}-\zeta)^2 \underline{\ell}_{\text{pred}}(\tilde{y})\mathrm{d}\tilde{y} \\
&\leq (1-\eta)\min_{\zeta\in\mathbb{R}} \int_{\mathbb{R}} (\tilde{y}-\zeta)^2 \ell_{\text{pred}}(\tilde{y})\mathrm{d}\tilde{y} \qquad (23) \\
&= (1-\eta)\mathbb{V}_{\ell_{\text{pred}}}(\tilde{Y}),
\end{aligned}$$

where (23) is an immediate consequence of Theorem 3, since for every fixed $\zeta \in \mathbb{R}$ the function $\tilde{y} \mapsto (\tilde{y}-\zeta)^2$ is nonnegative. □

Proof of Proposition 7. We have that

$$\begin{aligned}
\mathbb{V}_{\overline{\ell}_{\text{pred}}}(\tilde{Y}) &\coloneqq \min_{\zeta\in\mathbb{R}} \mathbb{E}_{\overline{\ell}_{\text{pred}}}[(\tilde{Y}-\zeta)^2] \\
&= \min_{\zeta\in\mathbb{R}} \int_{\mathbb{R}} (\tilde{y}-\zeta)^2 \overline{\ell}_{\text{pred}}(\tilde{y})\,\mathrm{d}\tilde{y} \\
&\geq \min_{\zeta\in\mathbb{R}} \int_{\mathbb{R}} (\tilde{y}-\zeta)^2 \left[(1-\eta)\ell_{\text{pred}}(\tilde{y}) + \eta u_{[a,b]}(\tilde{y})\right]\mathrm{d}\tilde{y} \qquad (24) \\
&\geq \min_{\zeta\in\mathbb{R}} \left[(1-\eta)\int_a^b (\tilde{y}-\zeta)^2 \ell_{\text{pred}}(\tilde{y})\,\mathrm{d}\tilde{y} + \eta\int_a^b (\tilde{y}-\zeta)^2 \frac{1}{b-a}\,\mathrm{d}\tilde{y}\right], \qquad (25)
\end{aligned}$$

where (24) follows from Corollary 3.1, and (25) follows by restricting the first integral to $[a,b]$, since the integrand is nonnegative.

By definition of the truncated density, for $\tilde{y}\in[a,b]$ we have

$$\ell_{\text{pred}}(\tilde{y}) = m\,\ell'_{\text{pred}}(\tilde{y}).$$

Therefore,

$$\begin{aligned}
\mathbb{V}_{\overline{\ell}_{\text{pred}}}(\tilde{Y}) &\geq \min_{\zeta\in\mathbb{R}} \left[(1-\eta)m\int_a^b (\tilde{y}-\zeta)^2 \ell'_{\text{pred}}(\tilde{y})\,\mathrm{d}\tilde{y} + \eta\int_a^b (\tilde{y}-\zeta)^2 \frac{1}{b-a}\,\mathrm{d}\tilde{y}\right] \\
&\geq (1-\eta)m \min_{\zeta\in\mathbb{R}} \int_a^b (\tilde{y}-\zeta)^2 \ell'_{\text{pred}}(\tilde{y})\,\mathrm{d}\tilde{y} + \eta \min_{\zeta\in\mathbb{R}} \int_a^b (\tilde{y}-\zeta)^2 \frac{1}{b-a}\,\mathrm{d}\tilde{y} \\
&= (1-\eta)m\,\mathbb{V}_{\ell'_{\text{pred}}}(\tilde{Y}) + \eta\frac{(b-a)^2}{12}. \qquad (26)
\end{aligned}$$

The last equality holds because the value $\zeta = \frac{a+b}{2}$ that minimizes

$$\int_a^b (\tilde{y}-\zeta)^2 \frac{1}{b-a}\,\mathrm{d}\tilde{y}$$

is the midpoint of the interval $[a,b]$, and the corresponding minimum is the variance of the uniform distribution on $[a,b]$, namely $(b-a)^2/12$.

Finally, if $m\approx 1$, that is, if $[a,b]$ captures most of the Gaussian mass of ℓ_{pred}, then (26) yields the approximation

$$\mathbb{V}_{\overline{\ell}_{\text{pred}}}(\tilde{Y}) \gtrsim (1-\eta)\mathbb{V}_{\ell'_{\text{pred}}}(\tilde{Y}) + \eta\frac{(b-a)^2}{12}.$$

□

We now present a lemma that will play a pivotal role in the proof of Lemma 8.

Lemma 9 (Predictive Variance Bounds: Sanity Check)**.** *It holds that* $\mathbb{V}_{\underline{\ell}_{pred}}(\tilde{Y}) \leq \mathbb{V}_{\ell_{pred}}(\tilde{Y}) \leq \mathbb{V}_{\overline{\ell}_{pred}}(\tilde{Y})$.

Proof of Lemma 9. By Walley [1991, Appendix G], we can write

$$\mathbb{V}_{\ell_{\text{pred}}}(\tilde{Y}) = \min_{\zeta\in\mathbb{R}} \mathbb{E}_{\ell_{\text{pred}}}[(\tilde{Y}-\zeta)^2].$$

Moreover, by Theorem 3,

$$\underline{\ell}_{\text{pred}}(\tilde{y}) \leq (1-\eta)\ell_{\text{pred}}(\tilde{y}) \leq \ell_{\text{pred}}(\tilde{y}),$$

and by Corollary 3.1, choosing $u = \ell_{\text{pred}}$, we obtain

$$\overline{\ell}_{\text{pred}}(\tilde{y}) \geq (1-\eta)\ell_{\text{pred}}(\tilde{y}) + \eta\ell_{\text{pred}}(\tilde{y}) = \ell_{\text{pred}}(\tilde{y}).$$

Hence, for each $\zeta\in\mathbb{R}$,

$$\mathbb{E}_{\underline{\ell}_{\text{pred}}}[(\tilde{Y}-\zeta)^2] \leq \mathbb{E}_{\ell_{\text{pred}}}[(\tilde{Y}-\zeta)^2] \leq \mathbb{E}_{\overline{\ell}_{\text{pred}}}[(\tilde{Y}-\zeta)^2],$$

because $(\tilde{y}-\zeta)^2 \geq 0$. Minimizing over ζ yields the claim. □

Proof of Lemma 8. The first inequality in (7) comes from Proposition 6, the second from η being in $[0, 1]$, the third from our assumption that $\frac{(b-a)^2}{12} \geq \mathbb{V}_{\ell_{\text{pred}}}(\tilde{Y})$, the following approximation from our assumption $\ell_{\text{pred}} \approx \ell'_{\text{pred}}$, and the fourth from Proposition 7 together with the assumption $m \approx 1$. □

Since the contamination bounds act as affine transformations of the baseline predictive variance, we first record a simple invariance property of extrema under positive affine transformations.

Lemma 10 (Affine Invariance of Extrema)**.** *Let $f : I \to \mathbb{R}$ be a real-valued function on an interval $I \subset \mathbb{R}$, and let $a > 0$, $b \in \mathbb{R}$. Define $g(x) \coloneqq af(x) + b$. Then,*

$$\arg\max_{x \in I} g(x) = \arg\max_{x \in I} f(x),$$
$$\arg\min_{x \in I} g(x) = \arg\min_{x \in I} f(x).$$

Proof. For any $x_1, x_2 \in I$,

$$f(x_1) \leq f(x_2) \quad \iff \quad af(x_1) + b \leq af(x_2) + b,$$

since $a > 0$. Thus the ordering of function values is preserved, and the locations of maxima and minima coincide. □

Proof of Corollary 8.1. Since $\eta < 1$, we have $1 - \eta > 0$, and therefore Lemma 10 applies to g_-. The function $g_-(\psi_1) = (1 - \eta)V_{\text{base}}(\psi_1; \lambda)$ is an affine transformation of V_{base} with positive multiplicative constant $1 - \eta$. By Lemma 10, g_- and V_{base} share the same maximizer(s).

Similarly, under the truncation approximation with

$$a(\psi_1) = \hat{f}(\tilde{\mathbf{x}}) + \alpha_0 s(\tilde{\mathbf{x}}), \qquad b(\psi_1) = \hat{f}(\tilde{\mathbf{x}}) + \beta_0 s(\tilde{\mathbf{x}}),$$

for fixed constants $\alpha_0 < \beta_0$ independent of ψ_1, the truncated Gaussian predictive variance satisfies

$$V'_{\text{base}}(\psi_1; \lambda) = c_{\alpha_0, \beta_0} V_{\text{base}}(\psi_1; \lambda)$$

for some constant $c_{\alpha_0, \beta_0} > 0$ depending only on α_0, β_0. Moreover,

$$\frac{(b-a)^2}{12} = \frac{(\beta_0 - \alpha_0)^2}{12} V_{\text{base}}(\psi_1; \lambda).$$

Hence

$$g_+(\psi_1) = \left((1 - \eta) c_{\alpha_0, \beta_0} + \eta \frac{(\beta_0 - \alpha_0)^2}{12} \right) V_{\text{base}}(\psi_1; \lambda),$$

which is a positive multiple of $V_{\text{base}}(\psi_1; \lambda)$. Applying Lemma 10 yields that g_+ and V_{base} have the same maximizer(s).

Therefore the bounding functions g_- and g_+ attain their maxima at $\psi_1^\star$, and form an envelope around the same double-descent peak. □