The Annals of Statistics
2019, Vol. 47, No. 1, 39–66
https://doi.org/10.1214/18-AOS1712

BAYESIAN FRACTIONAL POSTERIORS

BY ANIRBAN BHATTACHARYA[*,1], DEBDEEP PATI[*,2] AND YUN YANG[†]

Texas A&M University[*] and University of Illinois at Urbana–Champaign[†]

We consider the *fractional posterior* distribution that is obtained by updating a prior distribution via Bayes theorem with a *fractional likelihood* function, a usual likelihood function raised to a fractional power. First, we analyze the contraction property of the fractional posterior in a general misspecified framework. Our contraction results only require a prior mass condition on certain Kullback–Leibler (KL) neighborhood of the true parameter (or the KL divergence minimizer in the misspecified case), and obviate constructions of test functions and sieves commonly used in the literature for analyzing the contraction property of a regular posterior. We show through a counterexample that some condition controlling the complexity of the parameter space is necessary for the regular posterior to contract, rendering additional flexibility on the choice of the prior for the fractional posterior. Second, we derive a novel Bayesian oracle inequality based on a PAC-Bayes inequality in misspecified models. Our derivation reveals several advantages of averaging based Bayesian procedures over optimization based frequentist procedures. As an application of the Bayesian oracle inequality, we derive a sharp oracle inequality in multivariate convex regression problems. We also illustrate the theory in Gaussian process regression and density estimation problems.

1. Introduction and preliminaries. The usage of *fractional likelihoods* has generated renewed attention in Bayesian statistics in recent years, where one raises a likelihood function to a fractional power, and combines the resulting fractional likelihood with a prior distribution via the usual Bayes formula to arrive at a *power posterior* or *fractional posterior* distribution. Applications of fractional posteriors have been diverse, ranging from fractional Bayes factors in objective Bayesian model selection [48], data-dependent priors for sparse estimation [42, 43], to marginal likelihood approximation [18] and posterior simulation [19]. The fractional posteriors are a special instance of *Gibbs posteriors* [32] or *quasi-posteriors* [14], where the negative exponent of a loss function targeted toward a specific parameter of interest is used as a surrogate for the likelihood function; see [10] for a general framework for updating of prior beliefs using Gibbs posteriors.

Received November 2016; revised April 2018.

[1]Supported by NSF Grant DMS-1613156 and NSF CAREER Grant DMS-1653404.

[2]Supported by NSF Grant DMS-1613156.

MSC2010 subject classifications. Primary 62G07, 62G20; secondary 60K35.

Key words and phrases. Posterior contraction, Rényi divergence, misspecified models, PAC-Bayes, oracle inequality, convex regression.

The recent surge of interest in fractional posteriors can be largely attributed to its empirically demonstrated robustness to misspecification [25, 46]. For correctly specified, or well-specified (non)parametric models, there is now a rich body of literature [22, 23, 54] guaranteeing concentration of the posterior distribution around minimax neighborhoods of the true data generating distribution. However, susceptibility to model misspecification poses a potent concern even for Bayesian nonparametric models which aim to capture finer aspects of the data.

There is a comparatively smaller literature on large sample behavior of nonparametric Bayesian procedures under misspecification [17, 33, 50], where the general aim is to establish sufficient conditions under which the usual posterior distribution concentrates around the nearest Kullback–Leibler (KL) point to the truth inside the parameter space. However, these conditions are considerably more stringent than those in case of well-specified models, so that verification can be fairly nontrivial, along with comparatively limited scope of applicability. In fact, [26] empirically demonstrate through a detailed simulation study that even convergence to the nearest KL point may not take place in misspecified models. They instead recommend using a fractional posterior, with a data-driven approach to choose the fractional power; see also [25]. More recently, [46] proposed a coarsened posterior approach to combat model misspecification, where one conditions on neighborhoods of the empirical distribution rather than on the observed data while applying Bayes formula. When the neighborhood is defined based on the KL divergence, the coarsened posterior essentially is a fractional posterior.

These observations compel us to systematically study the concentration properties of fractional posteriors. Walker and Hjort [61] established consistency of power posteriors for well-specified models; see also [44] for rate results. Zhang [62] arrived at similar conclusions from a minimum complexity density estimation perspective. Jiang and Tanner [32] extended results of [62] to a Gibbs' posterior framework to deal with model misspecification in a high-dimensional classification problem. One of the main contributions of this article beyond the existing literature is a unified general treatment of misspecified and well-specified models through the introduction of a novel divergence measure, and the statistical implication of the theoretical results in terms of usage of heavy-tailed priors and shape-constrained estimation.

Specifically, we derive rates of convergence for the fractional posterior for general non-i.i.d. models in a misspecified model framework. The sufficient conditions for the fractional posterior to concentrate at the nearest KL point turn out to be substantially simpler compared to the existing literature on misspecified models. We state our concentration results for a novel class of Rényi-type divergence measures in a nonasymptotic environment, which in particular, imply Hellinger concentration in properly specified settings. The effect of flattening the likelihood shows up in the leading constant in the rate. The subexponential nature of the posterior tails allow us to additionally derive posterior moment bounds.

As one of our contributions, we show that the contraction rate of the fractional posterior is entirely determined by the prior mass assigned to appropriate KL neighborhoods of the true distribution, bypassing the construction of *sieves*[3] in the existing theory [3, 22, 33]. One practically important consequence is that concentration results can be established for the fractional posterior for a much broader class of priors compared to the regular posterior. We provide several examples on usage of heavy tailed hyperpriors in density estimation and regression, where the fractional posterior provably concentrates at a (near) minimax rate, while the regular posterior has inconclusive behavior. Another novel application of our result lies in shape constrained function estimation. Obtaining metric entropy estimates in such problems pose a stiff technical challenge and constitutes an active area of research [28]. The fractional posterior obviates the need to obtain such entropy estimates en route to deriving concentration bounds.

As a second contribution, we develop oracle inequalities for the fractional posterior based on a new PAC-Bayes inequality [11, 12, 15, 27, 45, 52] in a fully general Bayesian model. Many previous results on PAC-Bayes type inequalities are specifically tailored to classification (bounded loss, [11, 12, 52]) or regression (squared loss, [15, 27, 40, 52]) problems. Moreover, in the machine learning literature, a PAC-Bayes inequality is primarily used as a computational tool for controlling the generalization error by optimizing its upper bound over a restricted class of "posterior" distributions [11, 12]. There is a need to develop a general PAC-Bayes inequality and an accompanied general theory for analyzing the Bayesian risk that can be applied to a broader class of statistical problems. In this paper, we derive an oracle-type inequality for Bayesian procedures, which will be referred to as a *Bayesian oracle inequality* (BOI), based on a new PAC-Bayes inequality. Similar to the local Rademacher complexity [4] or local Gaussian complexity [5] in a frequenstist oracle inequality (FOI) for penalized empirical risk minimization procedures [34, 35], a BOI also involves a penalty term, which we refer to as *local Bayesian complexity*, that characterizes the local complexity of the parameter space. Roughly speaking, the local Bayesian complexity is defined as the inverse sample size times the negative logarithm of the prior mass assigned to certain Kullback–Leibler neighborhood around the (pseudo) true parameter. In the special case when the prior distribution is close to be "uniform" over the parameter space, the local Bayesian complexity becomes the inverse sample size times a local covering entropy, and our BOI recovers the convergence rates derived from local covering conditions [39]. Moreover, our BOI naturally leads to *sharp* oracle inequalities when the model is misspecified. For example, when applied to convex regression, we derive a sharp oracle inequality with minimax-optimal (up to $\log n$ factors) excess risk bound that extends the recent sharp oracle inequality obtained in [7] from dimension one to general dimensions $d \geq 1$ under suitable conditions.

[3]Compact subsets of the parameter space with a delicate balance between their size measured in terms of metric entropy and the prior probability of their complement.

Last but not the least, our analysis reveals several potential advantages of averaging based Bayesian procedures over optimization based frequentist procedures. First, due to the averaging nature of a Bayesian procedure, our averaging case analysis leading to a BOI is significantly simpler than a common worst case analysis leading to a FOI. For example, a local average type excess risk bound from a Bayesian procedure allows us to use simple probability tools, such as the Markov inequality and Chebyshev's inequality, to obtain a high probability bound for the excess risk, since the expectation operation exchanges with the averaging (integration) operation. This is different from a local supremum-type excess risk from a optimization procedure, where more sophisticated empirical process tools are exploited to obtain a high probability bound for excess risk [4, 41, 57, 59], due to the nonexchangeability between the expectation operation and the supremum operation. For further details about the comparison between BOI and FOI, please refer to Section 3.2. Second, a Bayesian procedure naturally leads to adaptation to unknown hyperparameters or tuning parameters. We show that by placing a hyperprior that distributes proper weights to different levels of the hyperparameter, a BOI adaptively leads to the optimal rate corresponding to the best choice of the hyperparameter.

We begin by introducing notation, and then review Rényi divergences as our key metric characterizing the contraction of fractional posteriors.

1.1. *Notation.* Let $C[0,1]^d$ and $C^\alpha[0,1]^d$ denote the space of continuous functions and the Hölder space of α-smooth functions $f : [0,1]^d \to \mathbb{R}$, respectively, endowed with the supremum norm $\|f\|_\infty = \sup_{t\in[0,1]^d} |f(t)|$. For $\alpha > 0$, the Hölder space $C^\alpha[0,1]^d$ consists of functions $f \in C[0,1]^d$ that have bounded mixed partial derivatives up to order $\lfloor\alpha\rfloor$, with the partial derivatives of order $\lfloor\alpha\rfloor$ being Lipschitz continuous of order $\alpha - \lfloor\alpha\rfloor$. Let $\|\cdot\|_1$ and $\|\cdot\|_2$, respectively, denote the L_1 and L_2 norm on $[0,1]^d$ with respect to the Lebesgue measure (i.e., the uniform distribution). To distinguish the L_2 norm with respect to the Lebesgue measure on $\mathbb{R}^d$, we use the notation $\|\cdot\|_{2,d}$. Throughout, C, C' denote positive constants whose value may change from one line to the other. For a finite set A, let $|A|$ denote the cardinality of A. The set of natural numbers is denoted by $\mathbb{N}$. $a \lesssim b$ denotes $a \le Cb$ for some constant $C > 0$. $J(\varepsilon, A, \rho)$ denotes the ε-covering number of the set A with respect to the metric ρ. The m-dimensional simplex is denoted by Δ^{m-1}. I_k stands for the $k \times k$ identity matrix. $\mathrm{N}_d(\mu, \Sigma)$ denotes the d-variate normal distribution with mean μ and covariance Σ, and $\mathcal{N}_d(z; m, \Sigma)$ its density evaluated at $z = (z_1, \ldots, z_d)^{\mathrm{T}}$.

1.2. *Rényi divergences.* Let P and Q be probability measures on a common probability space with a dominating measure μ, and let $p = dP/d\mu, q = dQ/d\mu$. The Hellinger distance $h^2(p,q) = (1/2)\int(\sqrt{p} - \sqrt{q})^2\, d\mu = 1 - A(p,q)$, where $A(p,q) = \int \sqrt{pq}\, d\mu$ denotes the Hellinger affinity. Let $D(p,q) =$

$\int p \log(p/q)\, d\mu$ denote the Kullback–Leibler (KL) divergence between p and q. For any $\alpha \in (0, 1)$, let

$$D_\alpha(p, q) = \frac{1}{\alpha - 1} \log \int p^\alpha q^{1-\alpha}\, d\mu \tag{1.1}$$

denote the Rényi divergence of order α. Let us also denote $A_\alpha(p, q) = \int p^\alpha q^{1-\alpha}\, d\mu = e^{-(1-\alpha)D_\alpha(p,q)}$, which we shall refer to as the α-affinity. When $\alpha = 1/2$, the α-affinity equals the Hellinger affinity. We recall some important inequalities relating the above quantities; additional details and proofs can be found in [58].

(R1) $0 \le A_\alpha(p, q) \le 1$ for any $\alpha \in (0, 1)$, which in particular implies that $D_\alpha(p, q) \ge 0$ for any $\alpha \in (0, 1)$.

(R2) $D_{1/2}(p, q) = -2 \log A(p, q) = -2 \log\{1 - h^2(p, q)\} \ge 2h^2(p, q)$ using the inequality $\log(1 + t) < t$ for $t > -1$.

(R3) For fixed p, q, $D_\alpha(p, q)$ is increasing in the order $\alpha \in (0, 1)$. Moreover, the following two-sided inequality shows the equivalence of D_α and D_β for $0 < \alpha \le \beta < 1$:

$$\frac{\alpha}{\beta} \frac{1 - \beta}{1 - \alpha} D_\beta \le D_\alpha \le D_\beta, \qquad 0 < \alpha \le \beta < 1.$$

(R4) By an application of L'Hospital's rule, $\lim_{\alpha \to 1_-} D_\alpha(p, q) = D(p, q)$.

The rest of the paper is organized as follows. Section 2 sets up the statistical background for the technical results. The main results of the paper are stated in Section 3, with contraction results in Section 3.1, and the PAC-Bayesian inequality and Bayesian oracle inequality in Section 3.2. Applications to well-specified and misspecified problems are discussed respectively in Section 4 and Section 5. We conclude with a discussion in Section 6. All proofs are deferred to the Supplementary Material [8].

2. Background. We will present our theory on the large sample properties of fractional posteriors in its full generality by allowing the model to be misspecified and the observations, denoted by $X^{(n)} = (X_1, X_2, \ldots, X_n)$, to be neither identically nor independently distributed (abbreviated as non-i.i.d.) [23]. Our result for non-i.i.d. observations can be applied to models with nonindependent observations such as Gaussian time series and Markov processes, or models with independent, nonidentically distributed (i.n.i.d.) observations such as Gaussian regression and density regression.

Specifically, we adopt the notation of [23] and let $(\mathcal{X}^{(n)}, \mathcal{A}^{(n)}, \mathbb{P}_\theta^{(n)} : \theta \in \Theta)$ be a sequence of statistical experiments with observations $X^{(n)}$, where θ is the parameter of interest in arbitrary parameter space Θ, and n is the sample size. For each θ, let $\mathbb{P}_\theta^{(n)}$ admit a density $p_\theta^{(n)}$ relative to a σ-finite measure $\mu^{(n)}$. Assume that $(x, \theta) \to p_\theta^{(n)}(x)$ is jointly measurable relative to $\mathcal{A}^{(n)} \otimes \mathcal{B}$, where $\mathcal{B}$ is a

σ-field on Θ. In Section 4 and Section 5, we consider examples where θ is a regression function subject to smoothness or shape constraints, or the density of the observations itself.

We place a prior distribution Π_n on $\theta \in \Theta$, and define the fractional likelihood of order $\alpha \in (0, 1)$ to be the usual likelihood raised to power α:

$$L_{n,\alpha}(\theta) = \big[p_\theta^{(n)}(X^{(n)})\big]^\alpha. \tag{2.1}$$

Let $\Pi_{n,\alpha}(\cdot)$ denote the posterior distribution obtained by combining the fractional likelihood $L_{n,\alpha}$ with the prior Π_n, that is, for any measurable set $B \in \mathcal{B}$,

$$\Pi_{n,\alpha}\big(B|X^{(n)}\big) = \frac{\int_B L_{n,\alpha}(\theta)\Pi_n(d\theta)}{\int_\Theta L_{n,\alpha}(\theta)\Pi_n(d\theta)} = \frac{\int_B e^{-\alpha r_n(\theta,\theta^\dagger)}\Pi_n(d\theta)}{\int_\Theta e^{-\alpha r_n(\theta,\theta^\dagger)}\Pi_n(d\theta)}, \tag{2.2}$$

where $r_n(\theta, \theta^\dagger) := \log\{p_{\theta^\dagger}^{(n)}(X^{(n)})/p_\theta^{(n)}(X^{(n)})\}$ is the negative log-likelihood ratio between θ and any other fixed parameter value $\theta^\dagger$. For example, we may choose $\theta^\dagger$ as the parameter θ_0 associated with the true data generating distribution, abbreviated as the true parameter. Clearly, $\Pi_{n,1}$ denotes the usual posterior distribution.

We allow the model to be misspecified by allowing θ_0 to lie outside the parameter space Θ. In misspecified models, the point θ^* in Θ that minimizes the KL divergence from $\mathbb{P}_{\theta_0}^{(n)}$, that is,

$$\theta^* := \operatorname*{argmin}_{\theta\in\Theta} D\big(p_{\theta_0}^{(n)}, p_\theta^{(n)}\big), \tag{2.3}$$

plays the role of θ_0 in well-specified models [33]. When the parameter space Θ is convex,[4] θ^* (if exists) is automatically unique (up to redefinition on a null-set of $p_{\theta_0}^{(n)}$). In some cases, a sufficient condition for uniqueness is identifiability under $p_{\theta_0}^{(n)}$; refer to the example of estimating densities using kernel mixtures in Section 3 of [33]. When genuine multiple minimum KL points occurs, [33] extended their theory of posterior contraction to a finite subset of these multiple points. Such extensions are possible for our results along similar lines, and hence not discussed further.

We introduce the divergence

$$D_{\theta_0,\alpha}^{(n)}(\theta, \theta^*) := \frac{1}{\alpha - 1}\log A_{\theta_0,\alpha}^{(n)}(\theta, \theta^*), \tag{2.4}$$

referred to as the α-divergence with respect to $\mathbb{P}_{\theta_0}^{(n)}$, or simply θ_0, to measure the closeness between any $\theta \in \Theta$ and θ^*, where

$$A_{\theta_0,\alpha}^{(n)}(\theta, \theta^*) := \int \left(\frac{p_\theta^{(n)}}{p_{\theta^*}^{(n)}}\right)^\alpha p_{\theta_0}^{(n)}\, d\mu^{(n)}$$

[4] In the sense of Lemma 2.1 below.

is an α-affinity between θ and θ^* with respect to θ_0. To the best of our knowledge, this divergence measure $D^{(n)}_{\theta_0,\alpha}(\theta,\theta^*)$ has not been previously used in the posterior concentration context, although the affinity $A^{(n)}_{\theta_0,\alpha}(\theta,\theta^*)$ briefly appears in [33].

REMARK. In the well-specified case where $\theta^* = \theta_0 \in \Theta$, $A^{(n)}_{\theta_0,\alpha}$ reduces to the usual α-affinity defined in Section 1.2, and $D^{(n)}_{\theta_0,\alpha}$ becomes the Rényi divergence of order α between $p^{(n)}_\theta$ and $p^{(n)}_{\theta_0}$:

$$D^{(n)}_\alpha(\theta,\theta_0) = D_\alpha\big(p^{(n)}_\theta, p^{(n)}_{\theta_0}\big) = \frac{1}{\alpha-1}\log\int \big\{p^{(n)}_\theta\big\}^\alpha \big\{p^{(n)}_{\theta_0}\big\}^{1-\alpha}\, d\mu^{(n)}. \tag{2.5}$$

Note we drop θ_0 from the subscript when $\theta^* = \theta_0$.

In general, $D^{(n)}_{\theta_0,\alpha}$ continues to define a divergence measure that satisfies $D^{(n)}_{\theta_0,\alpha}(\theta,\theta^*) \geq 0$ for $\theta \in \Theta$ and $D^{(n)}_{\theta_0,\alpha}(\theta^*,\theta^*) = 0$ in a variety of statistical problems. For example, in the normal means problem $Y \sim \mathrm{N}_n(\theta, \sigma^2 \mathrm{I}_n)$ with $\theta \in \mathbb{R}^n$, $D^{(n)}_{\theta_0,\alpha}$ defines a divergence measure if the parameter space for the *mean* θ is a closed convex set in $\mathbb{R}^n$; see equation (5.2) in Sectio 5.1 and Section S1 of SD for more details. The convexity condition is satisfied by a broad class of problems, including isotonic regression, and convex regression [13]. In the density estimation context, Lemma 2.1 below shows that $D^{(n)}_{\theta_0,\alpha}$ defines a divergence measure if the parameter space of *densities* is convex.

LEMMA 2.1 (Property of α-divergences). *If* $\{p^{(n)}_\theta : \theta \in \Theta\}$ *is convex*[5] *or* θ^* *is an interior point of* Θ, *then* $0 < A^{(n)}_{\theta_0,\alpha}(\theta,\theta^*) \leq 1$ *for any* $\alpha \in (0,1)$. *Therefore,* $D^{(n)}_{\theta_0,\alpha}$ *defines a divergence that satisfies* $D^{(n)}_{\theta_0,\alpha}(\theta,\theta^*) \geq 0$ *for* $\theta \in \Theta$ *and* $D^{(n)}_{\theta_0,\alpha}(\theta^*,\theta^*) = 0$.

When $\alpha \in (0,1)$, the proof of the lemma implies that $D^{(n)}_{\theta_0,\alpha}(\theta,\theta^*) = 0$ if and only if $p^{(n)}_\theta = p^{(n)}_{\theta^*}$ on the support of $\mathbb{P}^{(n)}_{\theta_0}$, since x^α is a strictly concave function on $[0,\infty)$. A concrete application of Lemma 2.1 to show nonnegativity of the divergence is provided in Section 5.2.

We will primarily focus on the following two cases in this paper.

Independent and identically distributed observations. When $X_1, X_2, \ldots, X_n$ are i.i.d. observations, $\mathbb{P}^{(n)}_\theta$ equals the n-fold product measure $\mathbb{P}^n_\theta := \bigotimes_{i=1}^n \mathbb{P}_\theta$, where $\mathbb{P}_\theta$ is the common distribution for the observations. $\mathcal{A}^{(n)}$ also takes a product form

[5] Given any $\theta, \theta' \in \Theta$, and $\omega \in (0,1)$, there exists $\bar\theta \in \Theta$ such that $p^{(n)}_{\bar\theta} = \omega p^{(n)}_\theta + (1-\omega) p^{(n)}_{\theta'}$.

as $\mathcal{A}^n := \bigotimes_{i=1}^n \mathcal{A}$, with $\mathcal{A}$ the common σ-field. The fractional likelihood function is

$$L_{n,\alpha}(\theta) = \prod_{i=1}^{n} \{p_\theta(X_i)\}^\alpha, \tag{2.6}$$

where p_θ is the common density indexed by $\theta \in \Theta$. The negative log-likelihood ratio $r_n(\theta, \theta^\dagger) = \sum_{i=1}^n \log\{p_{\theta^\dagger}(X_i)/p_\theta(X_i)\}$ becomes the sum of individual log density ratios. Moreover, the α-affinity and divergence can be simplified as $A^{(n)}_{\theta_0,\alpha}(\theta, \theta^*) = \{A_{\theta_0,\alpha}(\theta, \theta^*)\}^n$ and $D^{(n)}_{\theta_0,\alpha}(\theta, \theta^*) = nD_{\theta_0,\alpha}(\theta, \theta^*)$, where $A_{\theta_0,\alpha}(\theta, \theta^*)$ and $D_{\theta_0,\alpha}(\theta, \theta^*)$, respectively, are the α-affinity and divergence for $n = 1$.

Independent observations. In this case as well, $X_1, X_2, \ldots, X_n$ are independent observations. However, the ith observation X_i has an index-dependent distribution $\mathbb{P}_{\theta,i}$, which possesses a density $p_{\theta,i}$ relative to a σ-finite measure μ_i on $(\mathcal{X}_i, \mathcal{A}_i)$. Thus, we take the measure $\mathbb{P}^{(n)}_\theta$ equal to the product measure $\bigotimes_{i=1}^n \mathbb{P}_{\theta,i}$ on the product measurable space $\bigotimes_{i=1}^n (\mathcal{X}_i, \mathcal{A}_i)$. The fractional likelihood function takes a product form as

$$L_{n,\alpha}(\theta) = \prod_{i=1}^{n} \{p_{\theta,i}(X_i)\}^\alpha, \tag{2.7}$$

and the negative log-likelihood ratio $r_n(\theta, \theta^\dagger) = \sum_{i=1}^n \log\{p_{\theta^\dagger,i}(X_i)/p_{\theta,i}(X_i)\}$. The α-affinity and divergence can be decomposed, respectively, as $A^{(n)}_{\theta_0,\alpha}(\theta, \theta^*) = \prod_{i=1}^n A_{\theta_0,\alpha,i}(\theta, \theta^*)$ and $D^{(n)}_{\theta_0,\alpha}(\theta, \theta^*) = \sum_{i=1}^n D_{\theta_0,\alpha,i}(\theta, \theta^*)$, where $A_{\theta_0,\alpha,i}(\theta, \theta^*)$ and $D_{\theta_0,\alpha,i}(\theta, \theta^*)$ are the α-affinity and divergence associated with the ith observation X_i.

3. Contraction and Bayesian oracle inequalities for fractional posteriors. This section contains our main results. We discuss the contraction of fraction posteriors in Section 3.1, and present novel Bayesian oracle inequalities based on PAC-Bayes type bounds in Section 3.2.

3.1. *General concentration bounds*. In this subsection, we consider the asymptotic behavior of fractional posterior distributions and corresponding Bayes estimators based on non-i.i.d. observations under the general misspecified framework. We give general results on the rate of contraction of the fractional posterior measure towards the KL minimizer θ^* relative to the α-divergence $D^{(n)}_{\theta_0,\alpha}$.

For any $\theta^\dagger$, define a specific KL neighborhood of $\theta^\dagger$ with radius ε as

$$\begin{aligned} B_n(\theta^\dagger, \varepsilon; \theta_0) = \Big\{ \theta \in \Theta : \int p^{(n)}_{\theta_0} \log\big(p^{(n)}_{\theta^\dagger}/p^{(n)}_\theta\big)\, d\mu^{(n)} \le n\varepsilon^2, \\ \int p^{(n)}_{\theta_0} \log^2\big(p^{(n)}_{\theta^\dagger}/p^{(n)}_\theta\big)\, d\mu^{(n)} \le n\varepsilon^2 \Big\}. \end{aligned} \tag{3.1}$$

It is standard practice to make assumptions on the prior mass assigned to such KL neighborhoods to obtain the rate of posterior concentration in misspecified models [33].

REMARK. An alternative definition of $B_n(\theta^\dagger, \varepsilon; \theta_0)$ is obtained by replacing the second inequality in (3.1) by

$$\int p_{\theta_0}^{(n)} \log^2\bigl(p_{\theta^\dagger}^{(n)}/p_\theta^{(n)}\bigr)\,d\mu^{(n)} - \biggl[\int p_{\theta_0}^{(n)} \log\bigl(p_{\theta^\dagger}^{(n)}/p_\theta^{(n)}\bigr)\,d\mu^{(n)}\biggr]^2 \le n\varepsilon^2.$$

All subsequent theorems are valid with either definition of B_n.

With these notation, we present a *nonasymptotic* upper bound for the posterior probability assigned to complements of α-divergence neighborhoods of θ^* with respect to θ_0.

THEOREM 3.1 (Contraction of fractional posterior distributions). *Fix* $\alpha \in (0,1)$. *Recall* θ^* *from* (2.3). *Assume that* ε_n *satisfies* $n\varepsilon_n^2 \ge 2$ *and*

$$\Pi_n\bigl(B_n(\theta^*, \varepsilon_n; \theta_0)\bigr) \ge e^{-n\varepsilon_n^2}. \tag{3.2}$$

Then, for any $D \ge 2$ *and* $t > 0$,

$$\Pi_{n,\alpha}\biggl(\frac{1}{n}D_{\theta_0,\alpha}^{(n)}(\theta,\theta^*) \ge \frac{D+3t}{1-\alpha}\varepsilon_n^2 \Big| X^{(n)}\biggr) \le e^{-tn\varepsilon_n^2}$$

holds with $\mathbb{P}_{\theta_0}^{(n)}$ *probability at least* $1 - 2/\{(D-1+t)^2 n\varepsilon_n^2\}$.

Condition (3.2) appears routinely in theoretical analysis of regular posterior distributions, with numerous verifications available in the literature for well-specified models. For misspecified models, examples of verification of (3.2) in density estimation and regression problems can be found in Theorems 3.2 and 4.1 of [33]. We provide two new illustrations in Section 5.

Since the divergence $D_{\theta_0,\alpha}^{(n)}(\theta,\theta^*)$ reduces to the usual Rényi divergence in the well-specified case $\theta^* = \theta_0$, the contraction of the fractional posterior $\Pi_{n,\alpha}$ can be established for the entire family of divergence measures $\frac{1}{n}D_{\theta_0,\beta}^{(n)}(\theta,\theta_0)$; $\beta \in (0,1)$, using the equivalence of the Rényi divergences in (R3), with only a change in the leading constant multiplying the rate ε_n. This is true for all subsequent results in the well-specified case, and hence not discussed individually afterwards.

Theorem 3.1 characterizes the contraction of the fractional posterior measure where the posterior of $D_{\theta_0,\alpha}^{(n)}(\theta,\theta^*)$ exhibits a subexponentially decaying tail. As a direct consequence, we have the following corollary that characterizes the fractional posterior moments of $D_{\theta_0,\alpha}^{(n)}(\theta,\theta^*)$.

COROLLARY 3.2 (Fractional posterior moments). *Under the conditions of Theorem 3.1 we have that, for any* $k \geq 1$,

$$\int \left\{ \frac{1}{n} D^{(n)}_{\theta_0,\alpha}(\theta, \theta^*) \right\}^k \Pi_{n,\alpha}(d\theta | X^{(n)}) \leq \frac{C_1}{(1-\alpha)^k} \varepsilon_n^{2k},$$

holds with $\mathbb{P}^{(n)}_{\theta_0}$ *probability at least* $1 - C_2/\{n\varepsilon_n^2\}$, *where* (C_1, C_2) *are some positive constants depending on* k.

Implications for well-specified models. While Theorem 3.1 and Corollary 3.2 apply generally to the misspecified setting, it is instructive to first consider their implications in the well-specified setting, that is, when the data generating parameter $\theta_0 \in \Theta$. Setting $t = 1$ in Theorem 3.1 implies that the fractional posterior increasingly concentrates on ε_n-sized $D^{(n)}_{\theta_0,\alpha}$ neighborhoods of the true parameter θ_0. In particular, given (R2) and (R3), Theorem 3.1 implies that for any $\alpha \in (0, 1)$, the rate of concentration of the fractional posterior $\Pi_{n,\alpha}$ in the Hellinger metric is ε_n. Similar concentration results for the usual posterior distribution in the Hellinger metric were established in [22, 54] for the i.i.d. case, and in [23] for the non-i.i.d. case. Since the prior mass condition (3.2) appears as one of the sufficient conditions there as well, the fractional posterior achieves the same rate of concentration as the usual posterior (albeit up to constants) in all the examples considered in these works, which is typically minimax up to a logarithmic term for appropriately chosen priors. In addition to the prior mass condition (3.2), the sufficient conditions of [23] additionally require the construction of *sieves* $\mathcal{F}_n \subset \Theta$ whose ε_n-entropy in the Hellinger metric is stipulated to grow in the order $\lesssim e^{Cn\varepsilon_n^2}$, and at the same time, the prior probability assigned to the complement of the sieve is required to be exponentially small, that is, $\Pi_n(\mathcal{F}_n^c) \leq e^{-C' n\varepsilon_n^2}$. The existence of such sieves with suitable control over their metric entropy is a crucial ingredient of their theory, as it guarantees existence of exponentially consistent test functions [9, 38] to test the true density against complements of Hellinger neighborhoods of the form $\{\theta \in \mathcal{F}_n : h^2(p^{(n)}_\theta, p^{(n)}_{\theta_0}) \geq M\varepsilon_n^2\}$.

An important distinction for the fractional posterior in Theorem 3.1 is that the prior mass condition *alone* is sufficient to guarantee optimal concentration. This is important for at least two distinct reasons. First, the condition of exponentially decaying prior mass assigned to the complement of the sieve implies fairly strong restrictions on the prior tails and essentially rules out heavy-tailed prior distributions on hyperparameters. On the other hand, a much broader class of prior choices lead to provably optimal posterior behavior for the fractional posterior. Second, obtaining tight bounds on the metric entropy in nonregular parameter spaces, for example, in shape-constrained regression problems, can be a substantially nontrivial exercise [28], which is entirely circumvented using the fractional posterior approach. Specific examples of either kind are provided in Section 4.

While it may be argued that the conditions on the entropy and complement probability of the sieve are only sufficient conditions, a counterexample from [3] suggests that some control on the complexity of the parameter space is also necessary to ensure the consistency of a regular posterior when the model space is well specified. Specifically, in their example, the posterior tends to put all its mass on a set of distributions that are $\sqrt{2-\sqrt{2}}$ away from the true data generating distribution with respect to the Hellinger metric, even though the prior assigns positive probability over any ε-KL ball around the true parameter. As an implication, the fractional posterior can still achieve a certain rate of contraction for this problem even though the regular posterior is not consistent. In fact, the rate of concentration of the fractional posterior $\varepsilon_n = (1-\alpha)^{-1}n^{-1/3}$ for this problem, since their prior satisfies $\Pi_n(B_n(\theta_0, \varepsilon; \theta_0)) \geq e^{-C\varepsilon^{-1}}$ for some constant $C > 0$. Therefore, a combination of Theorem 3.1 and the counterexample in [3] shows that the fractional posterior has an *annealing effect* that can flatten the potential peculiar spikes in the regular posterior that are far away from the true parameter. However, this additional flexibility of the fractional posterior comes at a price—when the regular posterior contracts, then the α-fractional posterior will sacrifice a factor of $(1-\alpha)^{-1}$ in the rate of contraction.

The following theorem shows that for fixed n, the fractional posterior will almost surely converge to the regular posterior ($\alpha = 1$) as $\alpha \to 1_-$.

THEOREM 3.3 (Regular posterior as a limit of fractional posteriors). *For each n, we have*

$$\mathbb{P}_{\theta_0}^{(n)}\Big[\Pi_{n,1}\big(B|X^{(n)}\big) = \lim_{\alpha\to 1_-} \Pi_{n,\alpha}\big(B|X^{(n)}\big), \forall B \in \mathcal{B}\Big] = 1.$$

This theorem implies that although for a fixed $\alpha \in (0, 1)$, the fraction posterior has the annealing effect of flattening the posterior, it will eventually convergence to the regular posterior as $\alpha \to 1_-$ almost surely. This observation also justifies the empirical observation [21] that parallel tempering can boost the convergence of the posterior when the posterior contracts. However, when the posterior is ill-behaved—does not have consistency or has multimodality, then we need a very fine grid for the design of α as $\alpha \to 1_-$ in the parallel tempering algorithm, since otherwise all factional posteriors will only exhibit the one big mode around θ^* and miss the rest.

Implications for misspecified models. A key reference for Bayesian asymptotics in infinite-dimensional misspecified models is [33], where sufficient conditions analogous to the well-specified case were provided for the posterior to concentrate around θ^*. The primary technical difficulty in showing such a result compared to the well-specified case is the construction of test functions, for which [33] proposed a novel solution. Akin to the well-specified case for the regular posterior, the

sufficient conditions of [33] constitute of a prior thickness condition as in Theorem 3.1, and conditions on entropy numbers. However, the entropy number conditions (equations (2.2) and (2.5) in [33]) for the misspecified case are substantially harder to verify. In their Lemma 2.1, a simpler sufficient condition related their entropy number condition to ordinary entropy numbers. Further, in their Lemma 2.3, exploiting convexity of the parameter space, they established that the sufficient conditions of their Lemma 2.1 are satisfied by a weighted Hellinger distance

$$h_w^2\big(\theta^{(n)}, \theta^{*(n)}\big) = \frac{1}{4}\int \Big(\sqrt{p_{\theta^*}^{(n)}} - \sqrt{p_{\theta}^{(n)}}\Big)^2 \frac{p_{\theta_0}^{(n)}}{p_{\theta^*}^{(n)}}\, d\mu^{(n)},$$

which then amounts to obtaining entropy numbers in the weighted Hellinger metric. Such an exercise typically requires further assumptions on the behavior of $p_{\theta}^{(n)}/p_{\theta^*}^{(n)}$. For example, if $\sup_\theta |p_{\theta}^{(n)}/p_{\theta^*}^{(n)}|$ is finite, the ordinary Hellinger metric dominates the weighted Hellinger metric and it suffices to obtain covering numbers with respect to the ordinary Hellinger metric. Under this assumption, the authors proceeded to derive convergence rates for the regular posterior in a density estimation problem using Dirichlet process mixture priors. However, this assumption precludes the true density $p_{\theta_0}^{(n)}$ to have heavier tails than that prescribed by the model. For example, if the true density is heavier that the class of densities specified by the model, the assumption $\sup |p_{\theta}^{(n)}/p_{\theta^*}^{(n)}| < \infty$ is not satisfied. Typically, in the misspecified case, controlling the prior mass (3.2) in Theorem 3.1 requires certain tail conditions on $p_{\theta_0}^{(n)}$. However, Theorem 3.1 obviates the need to verify any entropy conditions for the fractional posterior. It thus avoids the need to assume $\sup |p_{\theta}^{(n)}/p_{\theta^*}^{(n)}| < \infty$, unless required to verify the prior mass condition.

For $\alpha = 1/2$, our divergence measure $D_{1/2}(\theta, \theta^*)$ dominates the weighted Hellinger distance in which [33] derive their convergence rate for the density estimation problem in Theorem 3.1. This can be readily seen from

$$\begin{aligned} 4h_w^2(\theta, \theta^*) &= 1 + \int \frac{p_{\theta}^{(n)}}{p_{\theta^*}^{(n)}} p_{\theta_0}^{(n)}\, d\mu^{(n)} - 2\int \bigg(\frac{p_{\theta}^{(n)}}{p_{\theta^*}^{(n)}}\bigg)^{1/2} p_{\theta_0}^{(n)}\, d\mu^{(n)} \\ &\le 2\bigg[1 - \int \bigg(\frac{p_{\theta}^{(n)}}{p_{\theta^*}^{(n)}}\bigg)^{1/2} p_{\theta_0}^{(n)}\, d\mu^{(n)}\bigg] \le D_{1/2}(\theta, \theta^*), \end{aligned}$$

where the last inequality follows from $\log x \le x - 1$ and the penultimate inequality follows from Lemma 2.1.

3.2. *PAC-Bayes bounds and Bayesian oracle inequalities*. In many problems, the performance of a (pseudo) Bayesian approach can be characterized via PAC-Bayes type inequalities [27, 45, 52]. Throughout the rest of the paper, we use p.m. as an abbreviation for probability measure. The form of PAC-Bayes inequality we

consider in this article takes the form as

$$\int R(\theta,\theta_0)\Pi_{n,\alpha}(d\theta|X^{(n)}) \le \int S_n(\theta,\theta_0)\rho(d\theta) + \frac{1}{\kappa_n}D(\rho,\Pi_n) + \mathrm{Rem} \qquad \forall \text{ p.m. } \rho \ll \Pi_n,$$

where R is a statistical risk function, κ_n is a tuning parameter, Rem is a remainder term and S_n is a function that measures the discrepancy between θ and θ_0 on the support of ρ. Typical PAC-Bayes inequalities [27, 45, 52] are slightly less general in the sense S_n is usually an empirical estimate of the risk function R calculated based on the training sample.

We present a PAC-Bayes inequality for the fractional posterior distribution, where the risk function R is a multiple of the α-Rényi divergence $D_\alpha^{(n)}$ in (2.5), and $S_n(\theta,\theta_0)$ a multiple of the negative log-likelihood ratio $r_n(\theta,\theta_0)$.

THEOREM 3.4 (PAC-Bayes inequality relative to θ_0). *Fix $\alpha \in (0,1)$. Then, for any $\varepsilon \in (0,1)$,*

$$\begin{aligned}\int\left\{\frac{1}{n}D_\alpha^{(n)}(\theta,\theta_0)\right\}\Pi_{n,\alpha}(d\theta|X^{(n)}) &\le \frac{\alpha}{n(1-\alpha)}\int r_n(\theta,\theta_0)\rho(d\theta) + \frac{1}{n(1-\alpha)}D(\rho,\Pi_n)\\ &\quad + \frac{1}{n(1-\alpha)}\log(1/\varepsilon),\end{aligned} \tag{3.3}$$

for all p.m. $\rho \ll \Pi_n$, with $\mathbb{P}_{\theta_0}^{(n)}$ probability at least $(1-\varepsilon)$.

Theorem 3.4 immediately implies an oracle type inequality for the Bayes estimator $\widehat{\theta}_B := \int_\Theta \theta\,\Pi_{n,\alpha}(d\theta|X^{(n)})$ whenever $D_\alpha^{(n)}(\cdot,\theta_0)$ is a convex function, via an application of Jensen's inequality,

$$\begin{aligned}\frac{1}{n}D_\alpha^{(n)}(\widehat{\theta}_B,\theta_0) &\le \frac{\alpha}{n(1-\alpha)}\int r_n(\theta,\theta_0)\rho(d\theta) + \frac{1}{n(1-\alpha)}D(\rho,\Pi_n)\\ &\quad + \frac{1}{n(1-\alpha)}\log(1/\varepsilon),\end{aligned} \tag{3.4}$$

for all probability measure $\rho \ll \Pi_n$. We call this inequality a *Bayesian oracle inequality*. The convexity condition is automatically satisfied in density estimation problems where the parameter θ is the density itself, using convexity of the Rényi divergence as a function of the density [58]. Another example is Gaussian regression, $p_\theta^{(n)} \equiv \mathrm{N}(\theta, \mathrm{I}_n)$, where $\theta \in \mathbb{R}^n$ denotes the mean. In this case, a direct calculation yields $D_\alpha^{(n)}(\theta,\theta_0) = (\alpha/2)\|\theta-\theta_0\|_2^2$ and the conclusion follows from the convexity of the squared Euclidean distance.

Let us compare the Bayesian oracle inequality (BOI) with frequentist oracle inequalities (FOI) [34, 35]. For convenience, we assume that the observations are i.i.d., and use $\mathbb{P}_n$ to represent the empirical measure $\frac{1}{n}\sum_{i=1}^n \delta_{X_i}$. For a function $f : \mathcal{X} \to \mathbb{R}$, define

$$\mathbb{P}_n f = \frac{1}{n}\sum_{i=1}^n f(X_i) \quad \text{and} \quad \mathbb{P}_{\theta_0} f = \mathbb{E}_{\theta_0} f(X). \tag{3.5}$$

Under this notation, a typical FOI takes a form as

$$\mathbb{P}_{\theta_0} f_{\widehat{\theta}} \le c \inf_{\theta\in\Theta} \mathbb{P}_{\theta_0} f_\theta + \Psi_n(r_n), \tag{3.6}$$

for some leading constant $c \ge 1$, where $\widehat{\theta}$ is the estimator of θ, for example, obtained by empirical risk minimization [6, 34]. Here, $\mathcal{F} = \{f_\theta : \mathcal{X} \to \mathbb{R}, \theta \in \Theta\}$ is a class of functions indexed by $\theta \in \Theta$, such as, a certain loss function $\ell(\cdot, X)$ evaluated at θ. The term $\inf_{\theta\in\Theta} \mathbb{P}_{\theta_0} f_\theta$ will be referred to as the approximation error term, reflecting the smallest loss incurred by approximating f_{θ_0} from $\mathcal{F}$. The second term $\Psi_n(r_n)$ in the display is an excess risk term that reflects certain local complexity measure of $\mathcal{F}$, such as the local Rademacher complexity [4] or local Gaussian complexity [5]. $\Psi_n(r_n)$ typically serves as a high probability upper bound to the supremum of the localized empirical process,

$$\sup_{\theta\in\Theta:\mathbb{P}_n f_\theta \le r_n} \{\mathbb{P}_n f_\theta - \mathbb{P}_{\theta_0} f_\theta\}, \tag{3.7}$$

up to some other remainder terms, where r_n is a critical radius obtained as the solution of an equation involving a certain function depending on Ψ_n.

Now let us look at the BOI (3.4), which can be rewritten as

$$\begin{aligned}
&\frac{1}{n} D_\alpha^{(n)}(\hat{\theta}_B, \theta_0) \\
&\quad \le \frac{\alpha}{n(1-\alpha)} \inf_{\theta\in\Theta} \mathbb{P}_{\theta_0} r_\theta + \frac{\alpha}{n(1-\alpha)} \int \{\mathbb{P}_n r_\theta - \mathbb{P}_{\theta_0} r_\theta\} \rho(d\theta) \\
&\qquad + \left\{ \frac{\alpha}{n(1-\alpha)} \int \left\{ \mathbb{P}_{\theta_0} r_\theta - \inf_{\theta\in\Theta} \mathbb{P}_{\theta_0} r_\theta \right\} \rho(d\theta) + \frac{1}{n(1-\alpha)} D(\rho, \Pi_n) \right. \\
&\qquad \left. + \frac{1}{n(1-\alpha)} \log(1/\varepsilon) \right\},
\end{aligned} \tag{3.8}$$

where $r_\theta(X) = \log\{p_{\theta_0}(X)/p_\theta(X)\}$ is the log density ratio. We observe that the first term on the right-hand side of (3.8) is the approximation error term, and the rest serves as the excess risk term. However, the excess risk term in BOI has two distinctions from that in FOI. First, different from the FOI that induces localization via either an iterative procedure [36] or solving the solution of an equation involving a certain function [4], a BOI induces localization via picking a measure ρ concentrating around the best approximation $\operatorname{argmin}_{\theta\in\Theta} \mathbb{P}_{\theta_0} r_\theta$ that balances between

the average approximation error $\int\{\mathbb{P}_{\theta_0}r_\theta - \inf_{\theta\in\Theta}\mathbb{P}_{\theta_0}r_\theta\}\rho(d\theta)$ and a penalty on the size of localization $D(\rho,\Pi_n)$. Second, in FOI the stochastic term characterizing the local complexity is based on a worse case analysis by taking the supremum as in (3.7), while BOI bounds the stochastic term based on an average case analysis via the average fluctuation

$$\int\{\mathbb{P}_n r_\theta - \mathbb{P}_{\theta_0}r_\theta\}\rho(d\theta).$$

Because we can exchange the expectation with integration, this local average form allows us to use simple probability tools, such as Markov's inequality and Chebyshev's inequality, to obtain bounds for the excess risk. This is different from the local supremum form (3.7), where expectation does not exchange with supremum, and we need much more sophisticated empirical process tools such as chaining and peeling techniques to bound the excess risk (see, e.g., [4, 41, 57, 59]).

As a simple illustration of applying Chebyshev's inequality to BOI or inequality (3.3) in Theorem 3.4 to obtain an explicit risk bound for the Bayes estimator, we present the following corollary. Recall the definition of the KL neighborhood $B_n(\theta_0,\varepsilon;\theta_0)$ defined in (3.1).

COROLLARY 3.5. *Suppose $\varepsilon\in(0,1)$ satisfies $n\varepsilon^2>2$ and $D>1$. With $\mathbb{P}_{\theta_0}^{(n)}$ probability at least $1-2/\{(D-1)^2n\varepsilon^2\}$,*

$$\begin{aligned}\int\left\{\frac{1}{n}D_\alpha^{(n)}(\theta,\theta_0)\right\}\Pi_{n,\alpha}\big(d\theta|X^{(n)}\big)\\ \le\frac{D\alpha}{1-\alpha}\varepsilon^2+\left\{-\frac{1}{n(1-\alpha)}\log\Pi_n\big(B_n(\theta_0,\varepsilon;\theta_0)\big)\right\}.\end{aligned}\tag{3.9}$$

In particular, if we let ε_n to be the Bayesian critical radius that is the smallest ε satisfies

$$-\frac{\log\Pi_n(B_n(\theta_0,\varepsilon;\theta_0))}{n\varepsilon}\le\varepsilon,$$

then with $\mathbb{P}_{\theta_0}^{(n)}$ probability at least $1-2/\{(D-1)^2n\varepsilon_n^2\}$,

$$\int\left\{\frac{1}{n}D_\alpha^{(n)}(\theta,\theta_0)\right\}\Pi_{n,\alpha}\big(d\theta|X^{(n)}\big)\le\frac{D\alpha+1}{1-\alpha}\varepsilon_n^2.$$

The main idea of the proof is to choose the probability measure ρ as $\Pi_n(\cdot\cap B_n(\theta_0,\varepsilon;\theta_0))/\Pi_n(B_n(\theta_0,\varepsilon;\theta_0))$; the restriction of the prior Π_n to $B_n(\theta_0,\varepsilon;\theta_0)$. Under this choice, we have $D(\rho,\Pi_n)=-\log\Pi_n(B_n(\theta_0,\varepsilon;\theta_0))$, and $\int r_n(\theta,\theta_0)\rho(d\theta)$ can be bounded by applying Chebyshev's inequality. If higher moment constraints on the likelihood ratio $r_n(\theta,\theta_0)$ is also included into the definition of $B_n(\theta_0,\varepsilon;\theta_0)$ in (3.1), then the probability bound for (3.9) to hold can be boosted (for details, see Section 2 in [24]).

According to Corollary 3.5, the overall risk bound in (3.9) is a balance between two terms: an approximation error term ε_n^2 and a local complexity measure term $-\frac{1}{n}\log \Pi_n(B_n(\theta_0, \varepsilon_n; \theta_0))$. For this reason, we will refer to the second term as the *local Bayesian complexity*. The local Bayesian complexity reflects the compatibility between the prior distribution and the parameter space: if Π_n is close to a uniform distribution over Θ, then $-\log \Pi_n(B_n(\theta_0, \varepsilon_n; \theta_0)) = \log\{1/\Pi_n(B_n(\theta_0, \varepsilon_n; \theta_0)\}$ is roughly the logarithm of the number of ε_n-balls needed to cover a neighborhood of θ_0 and, therefore, is related to the local covering entropy. On the other hand, if some prior knowledge about θ_0 is available, then we can combine these knowledge to increase the prior mass around θ_0, which may significantly boost the rate of convergence of the Bayes estimator. This observation is consistent with our previous intuition that averaging based (average case analysis) Bayesian approaches sometimes can be better than optimization based (worst case analysis) frequentist approaches. For example, when certain hyperparameter or tuning parameter, such as the regularity of a function class or sparsity level of a regression model, is unknown, then a Bayesian procedure naturally achieves adaptation to those unknown parameters by placing a prior on them that distributes proper weights to different levels of the hyperparameter (see our examples in Section 4). In contrast, a common way to select a tuning parameter in frequentist methods is via cross-validation or data-splitting. These approaches only use some proportion of data to do estimation, after learning the tuning parameter via the rest, which may not be the most efficient way to use data.

Although Theorem 3.4 is useful for obtaining a BOI, when transformed into form (3.6) the resulting leading constant c of the approximation error term in the BOI is typically strictly larger than 1, resulting in a nonsharp oracle inequality. Here, we call an oracle inequality sharp if the leading constant c in (3.6) is 1; see, for example, [16]. To solve this issue for the PAC-Bayes inequality in Theorem 3.4, we consider a second class of PAC-Bayes inequalities that directly characterizes the closeness between θ and the best approximation θ^* of θ_0 from Θ.

THEOREM 3.6 (PAC-Bayes inequality relative to θ^*). *Fix $\alpha \in (0,1)$. Then, for any $\varepsilon \in (0,1)$,*

$$\begin{aligned}&\int \left\{\frac{1}{n} D_{\theta_0,\alpha}^{(n)}(\theta, \theta^*)\right\} \Pi_{n,\alpha}(d\theta | X^{(n)}) \\ &\quad \le \frac{\alpha}{n(1-\alpha)} \int r_n(\theta, \theta^*)\rho(d\theta) + \frac{1}{n(1-\alpha)} D(\rho, \Pi_n) \\ &\qquad + \frac{1}{n(1-\alpha)} \log(1/\varepsilon),\end{aligned} \tag{3.10}$$

for all p.m. $\rho \ll \Pi_n$, with $\mathbb{P}_{\theta_0}^{(n)}$ probability at least $(1-\varepsilon)$.

Similar to Corollary 3.5 for a concrete Bayesian risk bound for characterizing the closeness between θ and θ_0, we have the following counterpart for θ and θ^*.

COROLLARY 3.7. *For any $\varepsilon \in (0,1)$ satisfying $n\varepsilon^2 > 2$ and $D > 1$, with $\mathbb{P}_{\theta_0}^{(n)}$ probability at least $1 - 2/\{(D-1)^2 n\varepsilon^2\}$,*

$$\int \left\{ \frac{1}{n} D_{\theta_0,\alpha}^{(n)}(\theta, \theta^*) \right\} \Pi_{n,\alpha}(d\theta | X^{(n)}) \leq \frac{D\alpha}{1-\alpha}\varepsilon^2 + \left\{ -\frac{1}{n(1-\alpha)} \log \Pi_n(B_n(\theta^*, \varepsilon; \theta_0)) \right\}. \tag{3.11}$$

In particular, if we let ε_n to be the Bayesian critical radius that is the smallest ε satisfies

$$-\frac{\log \Pi_n(B_n(\theta^*, \varepsilon; \theta_0))}{n\varepsilon} \leq \varepsilon,$$

then with $\mathbb{P}_{\theta_0}^{(n)}$ probability at least $1 - 2/\{(D-1)^2 n\varepsilon_n^2\}$,

$$\int \left\{ \frac{1}{n} D_{\theta_0,\alpha}^{(n)}(\theta, \theta^*) \right\} \Pi_{n,\alpha}(d\theta | X^{(n)}) \leq \frac{D\alpha + 1}{1-\alpha}\varepsilon_n^2.$$

We now illustrate how Corollary 3.7 leads to a sharp oracle inequality in the misspecified case (a concrete example is provided in Section 5.1). As noted previously, an oracle inequality is sharp in the misspecified case if the leading constant in front of the model space approximation term is 1, that is, $d(\hat{\theta}, \theta_0) \leq \inf_{\theta \in \Theta} d(\theta, \theta_0) + C\varepsilon_n^2$ for some distance metric $d(\cdot, \cdot)$. In statistical learning theory, the *regret* [49, 60] of an estimator is defined as $d(\hat{\theta}, \theta_0) - \inf_{\theta \in \Theta} d(\theta, \theta_0)$. A benchmark to compare regrets for different estimators is the *minimax regret* defined as $\min_{\hat{\theta}} \max_{\theta_0} [\mathbb{E}_{\theta_0}\{d(\hat{\theta}, \theta_0)\} - \inf_{\theta \in \Theta} d(\theta, \theta_0)]$. Regret bounds (misspecified case) are substantially harder to obtain compared to minimax risk bounds (well-specified case), and the rate of minimax regret can be different from the minimax risk [49]. Our general technique to derive a sharp oracle inequality for the Bayes estimator will imply that the Bayes estimator has minimax regret.

Suppose we are interested in certain metric d_n defined on the parameter space Θ, the square of which is weaker than the average α-divergence $\frac{1}{n} D_{\theta_0,\alpha}^{(n)}(\theta, \theta^*)$, that is,

$$\frac{1}{n} D_{\theta_0,\alpha}^{(n)}(\theta, \theta^*) \geq c_\alpha d_n^2(\theta, \theta^*), \qquad \theta \in \Theta,$$

where c_α is some positive constant that may depend on α. We assume that d_n^2 is convex in its first argument. For simplicity, we also assume that θ^* is also the minimizer of $d_n(\theta, \theta_0)$ over $\theta \in \Theta$. Under these assumptions, Corollary 3.7 along with the convexity assumption implies that with high probability, $\widehat{\theta}_B$ satisfies $d_n(\widehat{\theta}_B, \theta^*) \leq c'_\alpha \varepsilon_n$, where ε_n is the Bayesian critical radius. Now adding $d_n(\theta^*, \theta_0)$ to both sides of this inequality and applying the triangle inequality, we obtain

$$d_n(\widehat{\theta}_B, \theta_0) \leq \inf_{\theta \in \Theta} d_n(\theta, \theta_0) + c'_\alpha \varepsilon_n,$$

which is a sharp oracle inequality.

Sometimes, we may be interested in obtaining an oracle inequality for the squared loss d_n^2, when Θ is a vector space and d_n is induced by an inner product, denoted by $\langle\cdot,\cdot\rangle_n$. This is a more intricate problem, as the trivial bound $d_n^2(\widehat{\theta}_B,\theta_0) \le 2[d_n^2(\widehat{\theta}_B,\theta^*) + d_n^2(\theta^*,\theta_0)]$ renders the oracle inequality non-sharp. However, it is usually true when Θ is a convex set that

$$\frac{1}{n} D_{\theta_0,\alpha}^{(n)}(\theta,\theta^*) \ge c_\alpha\big(d_n^2(\theta,\theta^*) + 2\langle\theta-\theta^*,\theta^*-\theta_0\rangle_n\big) \qquad \forall \theta\in\Theta.$$

For example, this inequality holds for regression with fixed design, where d_n is the L_2 empirical norm (details can be found in Section 4.1). Again, by applying Corollary 3.7 and adding $d_n^2(\theta^*,\theta_0)$ to both sides of this inequality, we obtain

$$\begin{aligned} d_n^2(\widehat{\theta}_B,\theta_0) &= d_n^2(\widehat{\theta}_B,\theta^*) + 2\langle\widehat{\theta}_B-\theta^*,\theta^*-\theta_0\rangle_n + d_n^2(\theta^*,\theta_0) \\ &\le \inf_{\theta\in\Theta} d_n^2(\theta,\theta_0) + c'_\alpha\varepsilon_n^2, \end{aligned}$$

which is a sharp oracle inequality for the squared loss d_n^2. As an illustration of this technique, we derive a sharp oracle inequality for estimating a convex function in Theorem 5.1 when the true regression function is not necessarily convex.

Comparisons with previous work. The most relevant PAC-Bayes type result to ours, such as Theorem 3.4, is the Theorem 1 in [15], which focus on the regression setting $Y_i = f(x_i) + w_i$, where $\theta = f$ is the unknown regression function to be estimated, x_i's are the fixed design points and w_i's are the i.i.d. zero mean noise, corresponding to the i.n.i.d. observations. They propose to use the posterior mean of the following quasi-likelihood function as the estimator:

$$L_{n,\beta}(f) = \exp\left\{-\frac{1}{2\beta}\sum_{i=1}^n \big(Y_i - f(x_i)\big)^2\right\},$$

where according to their terms, β is a temperature parameter. In the special case when $w_i \sim N(0,\sigma^2)$ and $\beta = \sigma^2$, this function reduces to the likelihood function. They establish a PAC-Bayes inequality

$$\mathbb{E}_{\theta_0}^{(n)}\big[\|\hat{f} - f_0\|_n^2\big] \le \int \|f - f_0\|_n^2 \rho(df) + \frac{\beta}{n} D(\rho,\Pi_n) \qquad \forall \text{ p.m. } \rho \ll \Pi_n,$$

when $\beta \ge 4\sigma^2$, where $\hat{f}$ is the corresponding posterior mean. Therefore, their quasi-likelihood approach can be viewed as a special case under our fractional posterior with $\alpha \le 1/4$. Their proof is specialized to the empirical $L^2(\mathbb{P}_n)$ loss and requires the log-likelihood function to also take a sum of squares form. In contrast, our PAC-Bayes inequality generalizes the results in [15] to a more broader class of models. Moreover, the posterior expectation in $\int R(f,f_0)\Pi_{n,\alpha}(df|X^{(n)})$ in our PAC-Bayes inequality is taken outside the loss function R instead of plugging in the estimator as $R(\widehat{f}_B, f_0)$, which is always bounded above by $\int R(f,f_0)\Pi_{n,\alpha}(df|X^{(n)})$ when $R(f,f_0)$ is a convex function of f.

In the next two sections, we demonstrate the salient features of our theory through a number of illustrative examples.

4. Examples in the well-specified case. In this section, we focus on *well-specified models* and the next section considers examples under model-misspecification. We make use of Corollary 3.5 in this section and state the results in the form of PAC-Bayes bounds. We note that one could alternatively use Theorem 3.1 to obtain similar conclusions. We discuss three examples in this section, the first two in the context of Gaussian regression and the third in density estimation. Our first example considers shape-constrained estimation where the true function is assumed to be convex. We demonstrate the efficacy of the fractional posterior approach in delivering optimal concentration, where we bypass the need to compute the covering entropy of the convex function space. To best of our knowledge, such a result is not available in the Bayesian literature. The next two examples concern the classical Gaussian process regression and nonparametric density estimation problems, where we show that the fractional posterior optimally and adaptively concentrates at the true parameter value with substantially relaxed assumptions on the prior compared to existing theory. These two examples also theoretically justify the adaptive nature of Bayesian approaches.

4.1. *Nonparametric regression.* Consider the following nonparametric regression model with fixed design:

$$y_i = \mu(x_i) + \varepsilon_i, \qquad \varepsilon_i \sim \mathrm{N}\big(0, \sigma^2\big), i = 1, \ldots, n, \tag{4.1}$$

where $y_i \in \mathbb{R}$ is the response, $x_i \in [0,1]^d$ is the ith fixed design point, $\mu : [0,1]^d \to \mathbb{R}$ is the unknown regression function to be estimated and σ is the noise level. Given our general notation, the function μ plays the role of θ here, and $p_\mu^{(n)} \equiv \mathrm{N}_n(\widetilde{\mu}, \sigma^2 I_n)$, where $\widetilde{\mu} = (\mu(x_1), \ldots, \mu(x_n))^{\mathrm{T}}$.

To estimate μ, we place a prior Π over an appropriate function space $\mathcal{F}$. An examination of the fractional likelihood $L_{n,\alpha}(\mu)$ and the corresponding posterior $\Pi_{n,\alpha}(\mu)$ under (4.1) yields

$$\Pi_{n,\alpha}(\mu) = \frac{\{\mathcal{N}_n(y; \widetilde{\mu}, \sigma^2 \mathrm{I}_n)\}^\alpha \Pi(\mu)}{\int \{\mathcal{N}_n(y; \widetilde{\mu}, \sigma^2 \mathrm{I}_n)\}^\alpha \Pi(d\mu)} = \frac{\mathcal{N}_n(y; \widetilde{\mu}, \psi^2 \mathrm{I}_n)\Pi(\mu)}{\int \mathcal{N}_n(y; \widetilde{\mu}, \psi^2 \mathrm{I}_n)\Pi(d\mu)},$$

where $y = (y_1, \ldots, y_n)^{\mathrm{T}}$, and $\psi = \sigma/\sqrt{\alpha}$. Hence the fractional posterior for (4.1) is essentially a standard posterior with a different variance parameter in the likelihood. For simplicity, we henceforth assume that σ is known, and without loss of generality, equals one.

We use the notation $\|\cdot\|_{2,n}$ to denote the $L_2(\mathbb{P}_n)$ norm relative to the empirical measure $\mathbb{P}_n = n^{-1}\sum_{i=1}^n \delta_{x_i}$, and use $\langle \cdot, \cdot \rangle_n$ to denote the empirical inner product, that is, $\langle f, g \rangle_n = n^{-1}\sum_{i=1}^n f(x_i)g(x_i)$ for two functions f and g. Let μ_0 denote the true regression function, which we assume to be inside $\mathcal{F}$ in this section. The Rényi divergence between two multivariate Gaussian distributions with identity covariance is given by

$$D_\alpha^{(n)}(\mu, \mu_0) = \frac{n\alpha}{2}\|\mu - \mu_0\|_{2,n}^2,$$

which is proportional to the squared empirical $L_2(\mathbb{P}_n)$ norm between μ and μ_0. Hence, an application of Corollary 3.5 provides a risk bound for the $L_2(\mathbb{P}_n)$ norm in terms of the prior concentration $\log \Pi_n(B_n(\theta_0, \varepsilon; \theta_0))$. We provide two distinct examples below.

Convex regression. We first provide an illustration via convex regression, where $\mathcal{F}$ is a function space of a large class of d-dimensional convex functions over $[0,1]^d$ satisfying some minimal regularity conditions. Our fractional posterior framework becomes especially attractive in such problems, since it obviates the need to compute entropy numbers in restricted spaces, which can be a challenging exercise in itself [28].

It is recent practice in the frequentist literature to avoid additional smoothness assumptions on convex functions while studying rates of convergence [2, 29]. To that end, let $\partial\mu(x)$ denote the *sub-gradient* of the function μ at the point x, that is,

$$\partial\mu(x) = \{s \in \mathbb{R}^d : \mu(z) \geq \mu(x) + s^{\mathrm{T}}(z - x), \text{ for all } z \in [0,1]^d\}.$$

As in [2], define the class of convex, sub-differentiable, uniformly Lipschitz functions on $[0,1]^d$ as

$$\begin{aligned}\mathrm{Co}_L[0,1]^d = \{&\mu : [0,1]^d \to \mathbb{R}, \mu \text{ is convex},\\ &\|s\| \leq L \text{ for all } s \in \partial\mu(x), \partial\mu(x) \text{ is non empty for all } x\}.\end{aligned}$$

We model μ as a maximum of hyperplanes (which are always convex) [30], with a prior distribution for the number of affine functions over which the maximum is taken. Specifically, we let

$$\begin{aligned}\mu(x)|k, \{a_j^k, b_j^k\} &= \max_{j \in \{1,\ldots,k\}} [(a_j^k)^{\mathrm{T}} x + b_j^k],\\ \{(a_j^k)^{\mathrm{T}}, b_j^k\}^{\mathrm{T}}|k &\sim \mathrm{N}(0, \tau^2 \mathrm{I}_{d+1}), k \sim \pi_k.\end{aligned} \tag{4.2}$$

The following theorem shows that in the well-specified case where $\mu_0 \in \mathcal{F} \equiv \mathrm{Co}_L[0,1]^d$, with no additional smoothness condition on μ_0, we obtain a Bayes risk bound of the order $n^{-2/(4+d)}$ up to logarithmic terms, which coincides with the minimax risk under any $d \geq 1$ [2, 29].

THEOREM 4.1 (Bayesian risk in convex regression, well-specified case). *Consider the model* (4.1) *with* $\mu_0 \in \mathrm{Co}_L[0,1]^d$, *and the prior for* μ *satisfies* (4.2) *with* $\pi_k \geq \exp\{-Ck \log k\}$ *for some constant* $C > 0$, *then with* $\mathbb{P}_{\mu_0}^{(n)}$ *probability tending to one*,

$$\int \|\mu - \mu_0\|_{2,n}^2 \, d\Pi_{n,\alpha}(\mu) \leq \frac{C}{\alpha(1-\alpha)} \varepsilon_n^2, \tag{4.3}$$

where $\varepsilon_n = n^{-2/(4+d)} \log^t n$ *with* $t = 2/(4+d)$, *and* C *is some constant independent of* α.

Nonparametric GP regression. Next, we consider the regression model (4.1) with function space $\mathcal{F} = C[0,1]^d$. We assign to μ a Gaussian process prior with mean function $h_\mu : [0,1]^d \to \mathbb{R}$ and covariance kernel $c(x,x')$, a positive definite function from $[0,1]^d \times [0,1]^d \to \mathbb{R}$. We denote the prior by $\mu \sim \mathrm{GP}(h_\mu, c)$. We assume $h_\mu \equiv 0$ and work with the squared exponential covariance kernel $c_a(x,x') = e^{-a^2\|x-x'\|^2}$, with the prior for a satisfying

$$g(a) \geq A_1 a^p e^{-B_1 a^d \log^q a}. \tag{4.4}$$

With this assumption, we show that the fractional posterior concentrates at the minimax rate (up to logarithmic terms) adaptively over $\mu_0 \in C^\beta[0,1]^d$, where β is the unknown smoothness level of μ_0. To obtain the same result for the usual posterior, [56] require the prior on a to additionally satisfy an upper bound of the same order as the lower bound in (4.4), once again, ruling out heavy tailed priors.

THEOREM 4.2. *Consider the model* (4.1), *with a conditional GP prior* $\mu|a \sim \mathrm{GP}(0, c_a)$ *and suppose* $a \sim g(\cdot)$ *satisfies* (4.4). *If the true function* $\mu_0 \in C^\beta[0,1]^d$, *then* (4.3) *is satisfied with* $\varepsilon_n = n^{-\beta/(2\beta+d)}(\log n)^t$, *where* $t = \{(1+d) \vee q\}/(2 + d/\alpha)$.

Theorem 4.2 can be extended to other kernels in a straightforward manner.

4.2. *Nonparametric density estimation*. We assume $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} p$ and the goal is to estimate the unknown density p, so that p plays the role of θ here and $p_\theta^{(n)}(X^{(n)}) \equiv \prod_{i=1}^n p(X_i)$. We model the density p via a mixture of finite mixtures (MFM; [47]), which is a finite mixture model with a prior on the number of mixture components. With some minor modifications, the results can be adapted to infinite mixture models, such as Dirichlet process mixtures [37, 53]. Unlike existing literature [37, 53], our concentration results can accommodate *heavy tailed* prior distributions on the component specific means. Due to space constraints, the details are postponed to Section S3 of SD.

5. Examples in the misspecified case. We now present examples where the true parameter θ_0 lies outside Θ. We again state our results in the form of a PAC-Bayes inequality, now based on applying Corollary 3.7. As discussed in Section 2, we need to verify that $D^{(n)}_{\theta_0,\alpha}(\theta,\theta^*) \geq 0$ as this quantity is not guaranteed to be positive in general. In our first example, we revisit the convex regression problem from the last section where the true function need no longer be convex. We show by *direct verification* that $D^{(n)}_{\theta_0,\alpha}$ defines a valid divergence measure in this context. We then proceed to derive a sharp Bayesian oracle inequality that extends the recent sharp oracle inequality for one-dimensional convex regression obtained in [7] to *general dimension* $d \geq 1$. Our second example concerns density estimation where we apply Lemma 2.1 to guarantee positivity of the divergence and obtain rates of convergence using Corollary 3.7.

5.1. *Misspecified convex regression*. Due to space constraints, we have provided various details pertaining to this section in *Section* S1 *of the SD*.

Revisit the Gaussian regression model (4.1) with the parameter space $\mathcal{F} \equiv \mathrm{Co}_L[0,1]^d$ for the mean function. We continue to use the prior (4.2) for μ. The major difference of the analysis here from the previous section is that the true function μ_0 is no longer assumed to be in $\mathcal{F}$. Let $K := \{\widetilde{\mu} = (\mu(x_1), \ldots, \mu(x_n))^{\mathrm{T}} : \mu \in \mathcal{F}\} \subset \mathbb{R}^n$ denote the parameter space for the n-vector $\widetilde{\mu}$ created by evaluating the function μ at the design points.

First, a simple calculation (see SD) yields

$$D(p_{\mu_0}^{(n)}, p_{\mu}^{(n)}) = \frac{n}{2}\|\mu - \mu_0\|_{2,n}^2, \tag{5.1}$$

therefore, $\mu^* \in \mathcal{F}$ minimizing the KL divergence,

$$\mu^* := \operatorname*{argmin}_{\mu \in \mathcal{F}} \|\mu - \mu_0\|_{2,n}^2,$$

is a projection of μ onto $\mathcal{F}$ under the $L_2(\mathbb{P}_n)$ norm. Next, we show that the divergence measure $D_{\mu_0,\alpha}^{(n)}(\mu, \mu^*)$ is valid. To that end, some calculation (see SD) yields

$$\begin{aligned} D_{\mu_0,\alpha}^{(n)}(\mu, \mu^*) &= \frac{n\alpha}{2(1-\alpha)}\big[\|\mu - \mu_0\|_{2,n}^2 - \|\mu^* - \mu_0\|_{2,n}^2 - \alpha\|\mu - \mu^*\|_{2,n}^2\big] \\ &= \frac{T\alpha}{2(1-\alpha)}, \end{aligned}$$

where

$$T = \big[\|\widetilde{\mu} - \widetilde{\mu}_0\|^2 - \|\widetilde{\mu}^* - \widetilde{\mu}_0\|^2 - \alpha\|\widetilde{\mu} - \widetilde{\mu}^*\|^2\big].$$

Note that $\widetilde{\mu}^* = \mathrm{Proj}_K(\widetilde{\mu}_0)$, the Euclidean projection of $\widetilde{\mu}_0$ to the set K. Since K is a closed convex set in $\mathbb{R}^n$ (see SD), it is a standard fact from convex geometry (see, e.g., [51] and the SD) that the projection is uniquely defined and satisfies

$$\langle \widetilde{\mu} - \widetilde{\mu}^*, \widetilde{\mu}^* - \widetilde{\mu}_0 \rangle \geq 0 \qquad \forall \mu \in \mathcal{F}.$$

With some algebra, $T = (1-\alpha)\|\widetilde{\mu} - \widetilde{\mu}^*\|^2 + 2\langle \widetilde{\mu} - \widetilde{\mu}^*, \widetilde{\mu}^* - \widetilde{\mu}_0 \rangle \geq 0$ by the above inequality. This establishes the validity of the divergence.

To apply the PAC-Bayes inequality in Corollary 3.7, we need a handle on $B_n(\mu^*, \varepsilon; \mu_0)$. Here, we use the version of $B_n(\mu^*, \varepsilon; \mu_0)$ in the remark after equation (3.1). Some algebraic simplification (see SD) yields

$$B_n(\mu^*, \varepsilon; \mu_0) \supset \big\{\mu \in \mathcal{F} : \|\mu - \mu^*\|_{2,n}^2 + 2\langle \mu - \mu^*, \mu^* - \mu_0 \rangle_n \leq \varepsilon^2\big\}. \tag{5.2}$$

With these ingredients, we obtain the following *sharp* Bayesian oracle inequality, which generalizes the result of one-dimensional convex regression obtained in [7] to general dimension $d \geq 1$.

THEOREM 5.1 (Bayesian risk for convex regression, misspecified case). *Consider the model (4.1) with $\mu_0 \in C[0,1]^d$, and the prior for μ satisfying the conditions in Theorem 4.1. Let $K = n^{d/(4+d)}$, and suppose there exist $\bar{a}_j \in \mathbb{R}^d$ and $\bar{b}_j \in \mathbb{R}$ for $j = 1, \ldots, K$ such that the function $\bar{\mu}(x) := \max_{j\in\{1,\ldots,K\}}\{\bar{a}_j^{\mathrm{T}} x + \bar{b}_j\}$ satisfies*

$$\|\mu - \bar{\mu}\|_{2,n}^2 \le \|\mu - \mu^*\|_{2,n}^2 + \varepsilon_n^2/4$$

for any $\mu \in \mathrm{Co}_L[0,1]^d$, where ε_n is as in Theorem 4.1. Then with $\mathbb{P}_{\mu_0}^{(n)}$ probability tending to one,

$$\int \|\mu - \mu_0\|_{2,n}^2 \, d\Pi_{n,\alpha}(\mu) \le \inf_{\mu\in\mathrm{Co}_L[0,1]^d} \|\mu - \mu_0\|_{2,n}^2 + \frac{C}{\alpha(1-\alpha)}\varepsilon_n^2, \tag{5.3}$$

where C is some constant independent of α.

The assumption about $\bar{\mu}$ posits that the closest KL point μ^* can be well approximated by $\bar{\mu}$ in the model space which is a maxima of $K \asymp n\varepsilon_n^2$ many hyperplanes.

This sharp oracle inequality implies some geometric structure of the fractional posterior that cannot be obtained via a nonsharp one. Specifically, it can be shown that with large fractional posterior probability, $\mu - \mu^*$ is almost perpendicular to $\mu^* - \mu_0$; see SD for details.

5.2. *Misspecified density estimation.* Let $q_1, \ldots, q_K$ be densities supported on a compact set $\mathcal{X} \subset \mathbb{R}$. Let $\Delta^{K-1} = \{x \in \mathbb{R}^{K-1} : x_h \ge 0 \ \forall h, \sum_{h=1}^{K-1} x_i \le 1\}$ be the $(K-1)$-dimensional probability simplex, and for $\nu = (\nu_1, \ldots, \nu_{K-1}) \in \Delta^{K-1}$, let $\nu_K = 1 - \sum_{h=1}^{K-1} \nu_h$. Given i.i.d. data $X_1, \ldots, X_n \in \mathcal{X}$, model the common unknown density p as

$$p(x) := p_\nu(x) = \sum_{h=1}^{K} \nu_h q_h(x), \qquad \nu \in \Delta^{K-1}.$$

Examples of such a density model with fixed dictionary elements can be found in [55], Supplement. The parameter space $\Theta = \{p_\nu : \nu \in \Delta^{K-1}\}$. We consider a Dirichlet$(\alpha, \ldots, \alpha)$ prior on ν for some fixed $\alpha \le 1$, and denote the induced prior on $p \equiv p_\nu$ by Π.

When the true density $p_0 \notin \Theta$, the pseudo-true parameter $p^* = p_{\nu^*}$, with

$$\nu^* := \operatorname*{argmin}_{\nu\in\Delta^{K-1}} D(p_0, p_\nu) = \operatorname{argmax}_{\nu\in\Delta^{K-1}} \int_{\mathcal{X}} p_0(x)\log\big(p_\nu(x)\big)\,dx.$$

Since we are in an i.i.d. setup, $\frac{1}{n}D_{p_0,\alpha}^{(n)}(p, p^*) = D_{p_0,\alpha}(p, p^*)$. We show that $D_{p_0,\alpha}(\cdot,\cdot)$ defines a valid divergence measure by showing that Θ is convex and hence Lemma 2.1 applies. To see this, suppose $p_1, p_2 \in \Theta$. By definition, these exist $\nu_1, \nu_2 \in \Delta^{K-1}$ such that $p_1 = p_{\nu_1}$ and $p_2 = p_{\nu_2}$. For any $\omega \in (0,1)$, $(1-\omega)p_1 + \omega p_2 = p_{\bar{\nu}}$, where $\bar{\nu} = (1-\omega)\nu_1 + \omega\nu_2 \in \Delta^{K-1}$, and hence $p_{\bar{\nu}} \in \Theta$.

Note that ν^* may or may not be in the interior of Θ, and hence verifying the convexity condition is crucial. We illustrate this through two examples; details are in SD. First, consider $K = 2$, $\mathcal{X} = [0, 1]$, $q_1(x) = c_1(1+x)^{1/2}$, $q_2(x) = c_2(1+x)^{3/2}$ for $x \in [0, 1]$, and $p(x) = \nu_1 q_1(x) + \nu_2 q_2(x)$. Suppose the true density $p_0(x) = c_0(1+x)$ for $x \in [0, 1]$. Clearly, $p_0 \notin \Theta$, and

$$\int_{\mathcal{X}} p_0(x) \log\big(p_\nu(x)\big)\, dx = C + \log \nu_1 + \int_{\mathcal{X}} \log\bigg(1 + \frac{1-\nu_1}{\nu_1}\frac{q_2(x)}{q_1(x)}\bigg) p_0(x)\, dx,$$

where $C = \int_{\mathcal{X}} p_0(x)(\log q_1(x))\, dx$ does not depend on ν. The integral can be obtained in closed-form, and numerically maximizing the resulting expression produces a unique maxima at $\nu^* = (0.4870, 0.5130)$, which lies in the interior of the simplex. Hence, either condition of Lemma 2.1 is satisfied.

As a second example, continue to assume $K = 2$, $\mathcal{X} = [0, 1]$. Suppose $q_1(x) = c_1 e^{-x}$ and $q_2(x) = c_2 e^{-10x}$ for $x \in [0, 1]$. If the true density $p_0(x) = c_0 e^{-9x}$, the integral above can again be obtained in closed-form and numerically maximizing it leads to $\nu^* = (0, 1)$; see Figure S1 in SD. In this case, the (unique) minimizer lies on the boundary, but since the parameter space is convex, Lemma 2.1 still applies.

We are now prepared to state the concentration theorem.

THEOREM 5.2. *Suppose there exists constants $a, b > 0$ such that $a \le q_h(x) \le b$ for all $x \in \mathcal{X}$, $h = 1, \ldots, K$. Suppose also that the true density p_0 is bounded between a and b. Let $\varepsilon_n = \sqrt{K \log(n/K)/n}$ and $D > 1$. Then, with $p_0^{(n)}$ probability tending to one,*

$$\int D_{p_0,\alpha}(p, p^*)\Pi_{n,\alpha}\big(dp|X^{(n)}\big) \lesssim \frac{D\alpha + 1}{1-\alpha}\varepsilon_n^2.$$

Theorem 5.2 follows from an application of Corollary 3.7. A main ingredient is to show that

$$B_n\big(p^*, \varepsilon; p_0\big) \supset \bigg\{\nu \in \Delta^{K-1} : \sum_{h=1}^{K} |\nu_h - \nu_h^*| < \varepsilon^2\bigg\},$$

and the proof is completed by a standard small ball probability estimate for Dirichlet vectors.

6. Discussion. The study of concentration properties here complements the development of power posteriors from a coherent system of updating beliefs as discussed in [10]. When the model is misspecified, Bissiri, Holmes and Walker [10] argue that it is preferable to view $-\log p_\theta^{(n)}(X^{(n)})$ as a loss function relating the data $X^{(n)}$ with parameter θ. A formal Bayesian update combining a prior $\Pi(\cdot)$ with the above loss function necessarily has to take the form $\Pi_{n,\alpha}(\theta|X^{(n)}) \propto [p_\theta^{(n)}(X^{(n)})]^\alpha \Pi(\theta)$ to remain coherent. This coherence property assigns a special

place to power posteriors among the bigger class of Gibbs posteriors. The complementary view arising from the current presentation highlights the natural development of the divergence measure D_α from a power likelihood with exponent α, leading to a natural and useful metric for studying concentration in a large class of problems.

The coherence property of [10] holds for any $\alpha > 0$ and does not require α to be less than one. Indeed, there are instances in the literature where values of $\alpha > 1$ have been used. Holmes and Walker [31] provide an example of fitting a Poisson model to count data showing under-dispersion, where a power bigger than one is intuitively reasonable. A power larger than one also arises in a stochastic approximation of a complete likelihood in big data settings. If $X_1, \ldots, X_n$ are i.i.d. from a density p and $X'_1, \ldots, X'_m$ is a random subset of $\{X_1, \ldots, X_n\}$, then a stochastic approximation of the complete likelihood based on the random subset is $\prod_{i=1}^n p(X_i) \approx \prod_{j=1}^m [p(X'_j)]^{n/m}$, since $\sum_{i=1}^n \log p(X_i) \approx \frac{n}{m} \sum_{j=1}^m \log p(X'_j)$ by the strong law. Some other instances of a power larger than one from the machine learning literature include [1, 20]. However, for $\alpha > 1$, the D_α divergence is stronger than the Kullback–Leibler divergence and we do not expect the concentration results to hold in the stated generality for these stronger divergences. Second, D_α may cease to be convex for $\alpha > 1$, rendering statements about point estimates difficult. For these reasons, we have restricted attention to $\alpha < 1$ in the present paper.

Another issue beyond the scope of the paper is the choice of α. While it is reassuring that the rate of convergence remains unaffected by the choice of α, a principled procedure to choose α should reflect improved finite sample behavior. A detailed discussion regarding the optimal choice of α from a prediction perspective is available in [25]. For parametric models, [31] recently proposed an approach based on equating the prior expected gain in information between the prior and posterior from two experiments. It would be interesting to extend our concentration results with such data-driven choices of α.

SUPPLEMENTARY MATERIAL

Proofs of main results (DOI: 10.1214/18-AOS1712SUPP; .pdf). All proofs and additional details pertaining to Section 4 and Section 5 are provided in the supplementary document.

REFERENCES

[1] ALQUIER, P., RIDGWAY, J. and CHOPIN, N. (2016). On the properties of variational approximations of Gibbs posteriors. *J. Mach. Learn. Res.* **17** Paper No. 239, 41. MR3595173

[2] BALÁZS, G., GYÖRGY, A. and SZEPESVÁRI, C. (2015). Near-optimal max-affine estimators for convex regression. In *AISTATS*.

[3] BARRON, A., SCHERVISH, M. J. and WASSERMAN, L. (1999). The consistency of posterior distributions in nonparametric problems. *Ann. Statist.* **27** 536–561. MR1714718

[4] BARTLETT, P. L., BOUSQUET, O. and MENDELSON, S. (2005). Local Rademacher complexities. *Ann. Statist.* **33** 1497–1537. MR2166554
[5] BARTLETT, P. L. and MENDELSON, S. (2002). Rademacher and Gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.* **3** 463–482. MR1984026
[6] BARTLETT, P. L., MENDELSON, S. and PHILIPS, P. (2004). Local complexities for empirical risk minimization. In *Learning Theory. Lecture Notes in Computer Science* **3120** 270–284. Springer, Berlin. MR2177915
[7] BELLEC, P. C. and TSYBAKOV, A. B. (2015). Sharp oracle bounds for monotone and convex regression through aggregation. *J. Mach. Learn. Res.* **16** 1879–1892. MR3417801
[8] BHATTACHARYA, A., PATI, D. and YANG, Y. (2019). Supplement to "Bayesian fractional posteriors." DOI:10.1214/18-AOS1712SUPP.
[9] BIRGÉ, L. (1984). Sur un théorème de minimax et son application aux tests. *Probab. Math. Statist.* **3** 259–282. MR0764150
[10] BISSIRI, P. G., HOLMES, C. C. and WALKER, S. G. (2016). A general framework for updating belief distributions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 1103–1130. MR3557191
[11] CATONI, O. (2007). *Pac-Bayesian Supervised Classification*: *The Thermodynamics of Statistical Learning*. *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **56**. IMS, Beachwood, OH. MR2483528
[12] CATONI, O. and PICARD, J. (2004). *Statistical Learning Theory and Stochastic Optimization*: *Ecole D'Eté de Probabilités de Saint-Flour*, *XXXI*-2001. Springer, Berlin.
[13] CHATTERJEE, S. (2014). A new perspective on least squares under convex constraint. *Ann. Statist.* **42** 2340–2381. MR3269982
[14] CHERNOZHUKOV, V. and HONG, H. (2003). An MCMC approach to classical estimation. *J. Econometrics* **115** 293–346. MR1984779
[15] DALALYAN, A. and TSYBAKOV, A. B. (2008). Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Mach. Learn.* **72** 39–61.
[16] DALALYAN, A. S. and SALMON, J. (2012). Sharp oracle inequalities for aggregation of affine estimators. *Ann. Statist.* **40** 2327–2355. MR3059085
[17] DE BLASI, P. and WALKER, S. G. (2013). Bayesian asymptotics with misspecified models. *Statist. Sinica* **23** 169–187. MR3076163
[18] FRIEL, N. and PETTITT, A. N. (2008). Marginal likelihood estimation via power posteriors. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 589–607. MR2420416
[19] GELMAN, A. and MENG, X.-L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statist. Sci.* **13** 163–185. MR1647507
[20] GERMAIN, P., LACASSE, A., LAVIOLETTE, F. and MARCHAND, M. (2009). PAC-Bayesian learning of linear classifiers. In *Proceedings of the* 26*th Annual International Conference on Machine Learning* 353–360. ACM, New York.
[21] GEYER, C. J. and THOMPSON, E. A. (1995). Annealing Markov chain Monte Carlo with applications to ancestral inference. *J. Amer. Statist. Assoc.* **90** 909–920.
[22] GHOSAL, S., GHOSH, J. K. and VAN DER VAART, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28** 500–531. MR1790007
[23] GHOSAL, S. and VAN DER VAART, A. (2007). Convergence rates of posterior distributions for non-i.i.d. observations. *Ann. Statist.* **35** 192–223. MR2332274
[24] GHOSAL, S. and VAN DER VAART, A. (2007). Posterior convergence rates of Dirichlet mixtures at smooth densities. *Ann. Statist.* **35** 697–723. MR2336864
[25] GRÜNWALD, P. (2012). The safe Bayesian: Learning the learning rate via the mixability gap. In *Algorithmic Learning Theory. Lecture Notes in Computer Science* **7568** 169–183. Springer, Heidelberg. MR3042889
[26] GRÜNWALD, P. and VAN OMMEN, T. (2017). Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Anal.* **12** 1069–1103. MR3724979

[27] GUEDJ, B. and ALQUIER, P. (2013). PAC-Bayesian estimation and prediction in sparse additive models. *Electron. J. Stat.* **7** 264–291. MR3020421

[28] GUNTUBOYINA, A. and SEN, B. (2013). Covering numbers for convex functions. *IEEE Trans. Inform. Theory* **59** 1957–1965. MR3043776

[29] GUNTUBOYINA, A. and SEN, B. (2015). Global risk bounds and adaptation in univariate convex regression. *Probab. Theory Related Fields* **163** 379–411. MR3405621

[30] HANNAH, L. A. and DUNSON, D. B. (2013). Multivariate convex regression with adaptive partitioning. *J. Mach. Learn. Res.* **14** 3261–3294. MR3144462

[31] HOLMES, C. C. and WALKER, S. G. (2017). Assigning a value to a power likelihood in a general Bayesian model. *Biometrika* **104** 497–503. MR3698270

[32] JIANG, W. and TANNER, M. A. (2008). Gibbs posterior for variable selection in high-dimensional classification and data mining. *Ann. Statist.* **36** 2207–2231. MR2458185

[33] KLEIJN, B. J. K. and VAN DER VAART, A. W. (2006). Misspecification in infinite-dimensional Bayesian statistics. *Ann. Statist.* **34** 837–877. MR2283395

[34] KOLTCHINSKII, V. (2006). Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.* **34** 2593–2656. MR2329442

[35] KOLTCHINSKII, V. (2011). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems. Lecture Notes in Math.* **2033**. Springer, Heidelberg. MR2829871

[36] KOLTCHINSKII, V. and PANCHENKO, D. (2000). Rademacher processes and bounding the risk of function learning. In *High Dimensional Probability, II* (*Seattle, WA*, 1999). *Progress in Probability* **47** 443–457. Birkhäuser, Boston, MA. MR1857339

[37] KRUIJER, W., ROUSSEAU, J. and VAN DER VAART, A. (2010). Adaptive Bayesian density estimation with location-scale mixtures. *Electron. J. Stat.* **4** 1225–1257. MR2735885

[38] LE CAM, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer, New York. MR0856411

[39] LECAM, L. (1973). Convergence of estimates under dimensionality restrictions. *Ann. Statist.* **1** 38–53. MR0334381

[40] LEUNG, G. and BARRON, A. R. (2006). Information theory and mixing least-squares regressions. *IEEE Trans. Inform. Theory* **52** 3396–3410. MR2242356

[41] LUGOSI, G. and WEGKAMP, M. (2004). Complexity regularization via localized random penalties. *Ann. Statist.* **32** 1679–1697. MR2089138

[42] MARTIN, R., MESS, R. and WALKER, S. G. (2017). Empirical Bayes posterior concentration in sparse high-dimensional linear models. *Bernoulli* **23** 1822–1847. MR3624879

[43] MARTIN, R. and WALKER, S. G. (2014). Asymptotically minimax empirical Bayes estimation of a sparse normal mean vector. *Electron. J. Stat.* **8** 2188–2206. MR3273623

[44] MARTIN, R. and WALKER, S. G. (2016). Optimal Bayesian posterior concentration rates with empirical priors. Preprint. Available at arXiv:1604.05734.

[45] MCALLESTER, D. A. (1998). Some PAC-Bayesian theorems. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory* (*Madison, WI*, 1998) 230–234. ACM, New York. MR1811587

[46] MILLER, J. W. and DUNSON, D. B. (2015). Robust Bayesian inference via coarsening. Preprint. Available at arXiv:1506.06101.

[47] MILLER, J. W. and HARRISON, M. T. (2018). Mixture models with a prior on the number of components. *J. Amer. Statist. Assoc.* **113** 340–356. MR3803469

[48] O'HAGAN, A. (1995). Fractional Bayes factors for model comparison. *J. Roy. Statist. Soc. Ser. B* **57** 99–138. MR1325379

[49] RAKHLIN, A., SRIDHARAN, K. and TSYBAKOV, A. B. (2017). Empirical entropy, minimax regret and minimax risk. *Bernoulli* **23** 789–824. MR3606751

[50] RAMAMOORTHI, R. V., SRIRAM, K. and MARTIN, R. (2015). On posterior concentration in misspecified models. *Bayesian Anal.* **10** 759–789. MR3432239

[51] ROCKAFELLAR, R. T. (1997). *Convex Analysis*. Princeton Univ. Press, Princeton, NJ. MR1451876

[52] SHAWE-TAYLOR, J. and WILLIAMSON, R. C. (1997). A PAC analysis of a Bayesian estimator. In *Proceedings of the Tenth Annual Conference on Computational Learning Theory* 2–9. ACM, New York.

[53] SHEN, W., TOKDAR, S. T. and GHOSAL, S. (2013). Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. *Biometrika* **100** 623–640. MR3094441

[54] SHEN, X. and WASSERMAN, L. (2001). Rates of convergence of posterior distributions. *Ann. Statist.* **29** 687–714. MR1865337

[55] STEPHENS, M. (2016). False discovery rates: A new deal. *Biostatistics* **18** 275–294.

[56] VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2009). Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *Ann. Statist.* **37** 2655–2675. MR2541442

[57] VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*: *With Applications to Statistics*. Springer, New York. MR1385671

[58] VAN ERVEN, T. and HARREMOËS, P. (2014). Rényi divergence and Kullback–Leibler divergence. *IEEE Trans. Inform. Theory* **60** 3797–3820. MR3225930

[59] VAN DE GEER, S. (2000). *Empirical Processes in M-Estimation*. Cambridge Univ. Press, Cambridge.

[60] VAPNIK, V. N. and CHERVONENKIS, A. J. (1974). Theory of pattern recognition.

[61] WALKER, S. and HJORT, N. L. (2001). On Bayesian consistency. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **63** 811–821. MR1872068

[62] ZHANG, T. (2006). From ϵ-entropy to KL-entropy: Analysis of minimum information complexity density estimation. *Ann. Statist.* **34** 2180–2210. MR2291497

A. BHATTACHARYA
D. PATI
DEPARTMENT OF STATISTICS
TEXAS A&M UNIVERSITY
3143 TAMU
COLLEGE STATION, TEXAS 77843
USA
E-MAIL: anirbanb@stat.tamu.edu
debdeep@stat.tamu.edu

Y. YANG
DEPARTMENT OF STATISTICS
UNIVERSITY OF ILLINOIS AT URBANA–CHAMPAIGN
725 SOUTH WRIGHT STREET
CHAMPAIGN, ILLINOIS 61820
USA
E-MAIL: yyang@stat.fsu.edu