The Annals of Statistics
2013, Vol. 41, No. 5, 2359–2390
DOI: 10.1214/13-AOS1150

QUASI-BAYESIAN ANALYSIS OF NONPARAMETRIC INSTRUMENTAL VARIABLES MODELS[1]

BY KENGO KATO

University of Tokyo

This paper aims at developing a quasi-Bayesian analysis of the nonparametric instrumental variables model, with a focus on the asymptotic properties of quasi-posterior distributions. In this paper, instead of assuming a distributional assumption on the data generating process, we consider a quasi-likelihood induced from the conditional moment restriction, and put priors on the function-valued parameter. We call the resulting posterior quasi-posterior, which corresponds to "Gibbs posterior" in the literature. Here we focus on priors constructed on slowly growing finite-dimensional sieves. We derive rates of contraction and a nonparametric Bernstein–von Mises type result for the quasi-posterior distribution, and rates of convergence for the quasi-Bayes estimator defined by the posterior expectation. We show that, with priors suitably chosen, the quasi-posterior distribution (the quasi-Bayes estimator) attains the minimax optimal rate of contraction (convergence, resp.). These results greatly sharpen the previous related work.

1. Introduction.

1.1. *Overview.* Let (Y, X, W) be a triplet of scalar random variables, where Y is a dependent variable, X is an endogenous variable and W is an instrumental variable. Without loosing much generality, we assume that the support of (X, W) is contained in $[0, 1]^2$. The support of Y may be unbounded. We consider the nonparametric instrumental variables (NPIV) model of the form

$$\mathbb{E}[Y \mid W] = \mathbb{E}\big[g_0(X) \mid W\big], \tag{1}$$

where $g_0 : [0, 1] \to \mathbb{R}$ is an unknown structural function of interest. Alternatively, we can write the model in a more conventional form

$$Y = g_0(X) + U, \mathbb{E}[U \mid W] = 0,$$

where X is potentially correlated with U and hence $\mathbb{E}[U \mid X] \neq 0$.

A model of the form (1) is of principal importance in econometrics (see [28, 31]). From a statistical perspective, the problem of recovering the structural

Received December 2012; revised July 2013.

[1]Supported by the Grant-in-Aid for Young Scientists (B) (25780152) from the JSPS.

MSC2010 subject classifications. 62G08, 62G20.

Key words and phrases. Asymptotic normality, inverse problem, nonparametric instrumental variables model, quasi-Bayes, rates of contraction.

function g_0 is challenging since it is an *ill-posed* inverse problem with an additional difficulty of *unknown* operator [K in (2) ahead]. Statistical inverse problems, including the current problem, have attracted considerable interests in statistics and econometrics (see, e.g., [8, 9]). For mathematical background of inverse problems, we refer to [43].

To see that the problem of recovering the structural function g_0 is an ill-posed inverse problem, suppose that (X, W) has a square-integrable joint density $f_{X,W}(x, w)$ on $[0, 1]^2$ and denote by $f_W(w)$ the marginal density of W. Define the linear operator $K : L_2[0, 1] \to L_2[0, 1]$ by

$$(Kg)(w) = \mathbb{E}\big[g(X) \mid W = w\big] f_W(w) = \int g(x) f_{X,W}(x, w)\, dx.$$

Then the NPIV model (1) is equivalent to the operator equation

$$Kg_0 = h, \tag{2}$$

where $h(w) = \mathbb{E}[Y \mid W = w] f_W(w)$. Suppose that K is injective to guarantee identification of g_0.[2] The problem is that, even though K is injective, its inverse K^{-1} is not L_2-continuous since K is Hilbert–Schmidt (as $f_{X,W}(x, w)$ is square integrable on $[0, 1]^2$) and hence the lth largest singular value, denoted by κ_l, is approaching zero as $l \to \infty$ (see, e.g., [56]). In this sense, the problem of recovering g_0 from h is ill-posed.

Approaches to estimating the structural function g_0 are roughly classified into two types: the method involving the Tikhonov regularization [16, 28] and the sieve-based method [2, 5, 32, 45].[3] The minimax optimal rates of convergence in estimating g_0 are established in [11, 28], and they are achieved by the estimators proposed in [5, 28] under their respective assumptions. All the above mentioned studies are, however, from a purely frequentist perspective. Little is known about the theoretical properties of Bayes or quasi-Bayes analysis of the NPIV model. Exceptions are [18–20, 44].

This paper aims at developing a quasi-Bayesian analysis of the NPIV model, with a focus on the asymptotic properties of quasi-posterior distributions. The approach taken is quasi-Bayes in the sense that it neither needs to assume any specific distribution of (Y, X, W), nor has to put a nonparametric prior on the unknown likelihood function. The analysis is then based upon a quasi-likelihood induced from the conditional moment restriction. The quasi-likelihood is constructed by first estimating the conditional moment function $m(\cdot, g) = \mathbb{E}[Y - g(X) \mid W = \cdot]$

[2]This global identification condition is, however, not a trivial assumption; see the discussion after Assumption 2 in Section 3.2 as well as the last paragraph in the next subsection.

[3]The sieve-method is further classified into two types: the method using slowly growing finite-dimensional sieves with no or light penalties where the dimensions of sieves play the role of regularization, and the method using large-dimensional sieves with heavy penalties where the penalty terms play the role of regularization (see [10]).

in a nonparametric way, and taking $\exp\{-(1/2)\sum_{i=1}^{n}\hat{m}^2(W_i, g)\}$ as if it were a likelihood of g. For this quasi-likelihood, we put a prior on the function-valued parameter g. By doing so, formally, the posterior distribution for g may be defined, which we call "quasi-posterior distribution." This posterior corresponds to what [35] called "Gibbs posterior," and has a substantial interpretation (see Proposition 1 ahead). The quasi-Bayesian approach in this paper builds upon [12] where the dimension of the parameter of interest is finite and fixed.

We focus here on priors constructed on slowly growing finite-dimensional sieves (called "sieve or series priors"), where the dimensions of the sieve spaces (which grow with the sample size) play the role of regularization to deal with the problem of ill-posedness. Potentially, there are several choices in sieve spaces, but we choose to use wavelet bases to form sieve spaces. Wavelet bases are useful to treat smoothness function classes such as Hölder–Zygmund and Sobolev spaces in a unified and convenient way. We also use wavelet series estimation of the conditional moment function.[4]

Under this setup, we study the asymptotic properties of the quasi-posterior distribution. The results obtained are summarized as follows. First, we derive rates of contraction for the quasi-posterior distribution and establish conditions on priors under which the minimax optimal rate of contraction is attained. Here the contraction is stated in the standard L_2-norm. Second, we show asymptotic normality of the quasi-posterior of the first k_n generalized Fourier coefficients, where $k_n \to \infty$ is the dimension of the sieve space. This may be viewed as a nonparametric Bernstein–von Mises type result (see [54], Chapter 10, for the classical Bernstein–von Mises theorem for regular parametric models). Third, we derive rates of convergence of the quasi-Bayes estimator defined by the posterior expectation and show that under some conditions it attains the minimax optimal rate of convergence. Finally, we give some specific sieve priors for which the quasi-posterior distribution (the quasi-Bayes estimator) attains the minimax optimal rate of contraction (convergence, resp.). These results greatly sharpen the previous work of, for example, [44], as we will review below.

1.2. *Literature review and contributions.* Closely related are [20] and [44]. The former paper worked on the reduced form equation $Y = \mathbb{E}[g_0(X) \mid W] + V$ with $V = U + g_0(X) - \mathbb{E}[g_0(X) \mid W]$ and assumed V to be normally distributed. They considered a Gaussian prior on g, and the posterior distribution is also Gaussian (conditionally on the variance of V). They proposed to "regularize" the posterior and studied the asymptotic properties of the "regularized" posterior distribution and its expectation. Clearly, the present paper largely differs from [20] in that (i) we do not assume normality of the "error"; (ii) roughly speaking, Florens and

[4]This does not rule out the use of other bases such as the Fourier and Hermite polynomial bases. See Remark 3.

Simoni's method is tied with the Tikhonov regularization method, while ours is tied with the sieve-based method with slowly growing sieves. We note the settings of [18, 19] are largely different from the present paper; moreover in the NPIV example, some high-level conditions on estimated operators are assumed in [18, 19], and hence they are not directly comparable to the present paper. Liao and Jiang [44] developed an important unified framework in estimating conditional moment restriction models based on a quasi-Bayesian approach, and their scope is more general than ours. They analyzed NPIV models in detail in their Section 4. Their posterior construction is similar to ours such as the use of sieve priors, but differs from ours in detail. For example, [44] transformed the conditional moment restriction into unconditional moment restrictions with increasing number of restrictions. On the other hand, we directly work on the conditional moment restriction, although whether Liao and Jiang's approach will lose any efficiency in the frequentist sense is not formally clear.

Importantly and substantially, neither [20] nor [44] established sharp contraction rates for their (quasi-)posterior distributions, nor asymptotic normality results. It is unclear whether Florens and Simoni's [20] rates (in their Theorem 2) are optimal, since their assumptions are substantially different from the past literature such as [28] and [11]; moreover, strictly speaking [20] did not formally derive contraction rates for their regularized posterior when the operator is unknown (note that [18, 19], though not directly comparable to the present paper, also did not formally derive posterior contraction rates in the NPIV example). Liao and Jiang [44] only established posterior consistency. Here we focus on a simple but important model, and establish the sharper asymptotic results for the quasi-posterior distribution. Notably, a wide class of (finite dimensional) sieve priors is shown to lead to the optimal contraction rate. Moreover, in [44], a point estimator of the structural function is not formally analyzed. Hence, the primal contribution of this paper is to considerably deepen the understanding of the asymptotic properties of the quasi-Bayesian procedure for the NPIV model.

The present paper deals with a quasi-Bayesian analysis of an infinite-dimensional model. The literature on theoretical studies of Bayesian analysis of infinite-dimensional models is large. See [24–27, 38, 50] for general contraction rates results for posterior distributions in infinite-dimensional models. Note that these results do not directly apply to our case: the proof of the main general theorem (Theorem 1) depends on the construction of suitable "tests" (see the proof of Proposition 4), but how to construct such tests in a specific problem in a nonlikelihood framework is not trivial, especially in the current NPIV model where we have to deal with the ill-posedness of inverse problem. Moreover, Proposition 4 alone is not sufficient for obtaining sharp contraction rates and an additional work is needed (see the proof of Theorem 1).

There is also a large literature on the Bayesian analysis of (ill-posed) inverse problems. One stream of research on this topic lies in the applied mathematics literature; see [51] and references therein. However, their models and scopes are substantially different from those of the present paper; for example, [29, 30] considered (ill-conditioned) finite-dimensional linear regression models with Gaussian errors and priors, and contractions rates of posterior distributions are not formally studied there. In the statistics literature, we may refer to [1, 15, 39–41] (in addition to [18–20, 44] that are already discussed), although their results are not applicable to the analysis of NPIV models because of its particular structure (i.e., especially the operator K is unknown, and non-Gaussian "errors" and priors are allowed). Hence the present paper provides a further contribution to the Bayesian analysis of ill-posed inverse problems.

Our asymptotic normality result builds upon the previous work on asymptotic normality of (quasi-)posterior distributions for models with increasing number of parameters [3, 4, 6, 7, 13, 22, 23]. Related is [6], in which the author established Bernstein–von Mises theorems for Gaussian regression models with increasing number of regressors and improved upon the earlier work of [22] in several aspects. Reference [6] covered nonparametric models by taking into account modeling bias in the analysis. However, none of these papers covered the NPIV model, nor more generally linear inverse problems.

Finally, while we here assume injectivity of the operator K in (2), as one of anonymous referees pointed out, this condition is not a trivial assumption (see also the discussion after Assumption 2 in Section 3.2), and there are a number of works that relax the injectivity assumption and explore partial identification approach, such as [42, 44, 46] and [10], Appendix A.

1.3. *Organization and notation.* The remainder of the paper is organized as follows. Section 2 gives an informal discussion of the quasi-Bayesian analysis of the NPIV model. Section 3 contains the main results of the paper where general theorems on contraction rates and asymptotic normality for quasi-posterior distributions, as well as convergence rates for quasi-Bayes estimators, are stated. Section 4 analyzes some specific sieve priors. Section 5 contains the proofs of the main results. Section 6 concludes with some further discussions. The Appendix contains some omitted technical results. Because of the space limitation, the Appendix is contained in the supplemental file [36].

Notation: For any given (random or nonrandom, scalar or vector) sequence $\{z_i\}_{i=1}^n$, we use the notation $\mathbb{E}_n[z_i] = n^{-1}\sum_{i=1}^n z_i$, which should be distinguished from the population expectation $\mathbb{E}[\cdot]$. For any vector z, let $z^{\otimes 2} = zz^T$ where z^T is the transpose of z. For any two sequences of positive constants r_n and s_n, we write $r_n \lesssim s_n$ if the ratio r_n/s_n is bounded, and $r_n \sim s_n$ if $r_n \lesssim s_n$ and $s_n \lesssim r_n$. Let $L_2[0,1]$ denote the usual L_2 space with respect to the Lebesgue measure for functions defined on $[0,1]$. Let $\|\cdot\|$ denote the L_2-norm, that is, $\|f\|^2 = \int_0^1 f^2(x)\,dx$. The inner product in $L_2[0,1]$ is denoted by $\langle\cdot,\cdot\rangle$, that is, $\langle f, g\rangle = \int_0^1 f(x)g(x)\,dx$. Let $C[0,1]$ denote the metric space of all continuous functions on $[0,1]$, equipped with the uniform metric. The Euclidean norm is denoted by $\|\cdot\|_{\ell^2}$. For any matrix A, let $s_{\min}(A)$ and $s_{\max}(A)$ denote the minimum and maximum singular values of A, respectively. Let $\|A\|_{\mathrm{op}}$ denote the operator norm of a matrix A [i.e., $\|A\|_{\mathrm{op}} = s_{\max}(A)$]. Denote by $dN(\mu, \Sigma)(x)$ the density of the multivariate normal distribution with mean vector μ and covariance matrix Σ.

2. Quasi-Bayesian analysis: Informal discussion. In this section, we outline a quasi-Bayesian analysis of the NPIV model (1). The discussion here is informal. The formal discussion is given in Section 3.

Let $\mathcal{G}$ be a parameter space (say, some smoothness class of functions, such as a Hölder–Zygmund or Sobolev space), for which we assume $g_0 \in \mathcal{G}$. We assume that $\mathcal{G}$ is at least contained in $C[0,1]$: $\mathcal{G} \subset C[0,1]$. Define the conditional moment function as $m(W, g) = \mathbb{E}[Y - g(X) \mid W]$, $g \in \mathcal{G}$. Then g_0 satisfies the conditional moment restriction

$$m(W, g_0) = 0, \qquad \text{a.s.} \tag{3}$$

Equivalently, we have $\mathbb{E}[m^2(W, g_0)] = 0$.

In this paper, for the purpose of robustness, any specific distribution of (Y, X, W) is not assumed, which we believe is more practical in statistical and econometric applications. So a Bayesian analysis in the standard sense is not applicable here since a proper likelihood for g (g is a generic version of g_0) is not available. Instead, we use a quasi-likelihood induced from the conditional moment restriction (3).

Let $(Y_1, X_1, W_1), \dots, (Y_n, X_n, W_n)$ be i.i.d. observations of (Y, X, W). Let $W^n = \{W_1, \dots, W_n\}$ and $\mathcal{D}_n = \{(Y_1, X_1, W_1), \dots, (Y_n, X_n, W_n)\}$. By (3), a plausible candidate of the quasi-likelihood would be

$$p_g(W^n) = \exp\{-(n/2)\mathbb{E}_n[m^2(W_i, g)]\},$$

since $p_g(W^n)$ is maximized at the true structural function g_0. However, this $p_g(W^n)$ is infeasible since $m(\cdot, g)$ is unknown. Instead of using $p_g(W^n)$, we replace $m(\cdot, g)$ by a suitable estimate $\hat{m}(\cdot, g)$ and use the quasi-likelihood of the form

$$p_g(\mathcal{D}_n) = \exp\{-(n/2)\mathbb{E}_n[\hat{m}^2(W_i, g)]\}.$$

Below we use a wavelet series estimator of $m(\cdot, g)$.

The quasi-Bayesian analysis considered here uses this quasi-likelihood as if it were a proper likelihood and puts priors on $g \in \mathcal{G}$. In this paper, as in [44], we shall use sieve priors (more precisely, priors constructed on slowly growing sieves; [44] indeed considered another class of priors, see their supplementary material). The basic idea is to construct a sequence of finite-dimensional sieves (say, $\mathcal{G}_n$) that well approximates the parameter space $\mathcal{G}$ (i.e., each function in $\mathcal{G}$ is well approximated by some function in $\mathcal{G}_n$ as n becomes large), and put priors concentrating on these sieves. Each sieve space is a subset of a linear space spanned by some basis functions. Hence the problem reduces to putting priors on the coefficients on those basis functions. Such priors are typically called "(finite dimensional) sieve priors" (or "series priors") and have been widely used in the nonparametric Bayesian and quasi-Bayesian analysis (see, e.g., [24, 25, 48]).

Let Π_n be a so-constructed prior on $g \in \mathcal{G}$. Then, formally, the posterior-like distribution of g given $\mathcal{D}_n$ may be defined by

$$\Pi_n(dg \mid \mathcal{D}_n) = \frac{p_g(\mathcal{D}_n)\Pi_n(dg)}{\int p_g(\mathcal{D}_n)\Pi_n(dg)}, \tag{4}$$

which we call "quasi-posterior distribution." The quasi-posterior distribution is not a proper posterior distribution in the strict Bayesian sense since $p_g(\mathcal{D}_n)$ is not a proper likelihood. Nevertheless, $\Pi_n(dg \mid \mathcal{D}_n)$ is a proper distribution, that is, $\int \Pi_n(dg \mid \mathcal{D}_n) = 1$. Similar to proper posterior distributions, contraction of the quasi-posterior distribution around g_0 intuitively means that it contains more and more accurate information about the true structural function g_0 as the sample size increases. Hence, as in proper posterior distributions, it is of fundamental importance to study rates of contraction of quasi-posterior distributions. Here we say that the quasi-posterior $\Pi_n(dg \mid \mathcal{D}_n)$ contracts around g_0 at rate $\varepsilon_n \to 0$ if $\Pi_n(g : \|g - g_0\| > \varepsilon_n \mid \mathcal{D}_n) \xrightarrow{P} 0$.

This quasi-posterior corresponds to what [58] called "Gibbs algorithm" and what [35] called "Gibbs posterior." The framework of the quasi-posterior (Gibbs posterior) allows us a flexibility since a stringent distributional assumption, such as normality, on the data generating process is not required. Such a framework widens a Bayesian approach to broad fields of statistical problems.[5] Moreover, the following proposition gives an interesting interpretation of the quasi-posterior.

PROPOSITION 1. *Let $\eta > 0$ be a fixed constant. Let Π be a prior distribution for g defined on, say, the Borel σ-field of $C[0,1]$. Suppose that the data $\mathcal{D}_n$ are fixed and the maps $g \mapsto \hat{m}_i(W_i, g)$ are measurable with respect to the Borel σ-field of $C[0,1]$. Then, the distribution*

$$\hat{\Pi}_\eta(dg) = \frac{\exp(-\eta \sum_{i=1}^n \hat{m}^2(W_i, g))\Pi(dg)}{\int \exp(-\eta \sum_{i=1}^n \hat{m}^2(W_i, g))\Pi(dg)}$$

minimizes the empirical information complexity defined by

$$\check{\Pi} \mapsto \int \sum_{i=1}^n \hat{m}^2(W_i, g)\check{\Pi}(dg) + \eta^{-1} D_{\mathrm{KL}}(\check{\Pi} \parallel \Pi) \tag{5}$$

[5] Jiang and Tanner ([35], page 2211) remarked: "This framework of the Gibbs posterior has been overlooked by most statisticians for a long time [$\cdots$] a foundation for understanding the statistical behavior of the Gibbs posterior, which we believe will open a productive new line of research."

over all distributions $\check{\Pi}$ *absolutely continuous with respect to* Π. *Here*

$$D_{\mathrm{KL}}(\check{\Pi} \,\|\, \Pi) = \int \check{\pi} \log \check{\pi}\, \Pi(dg) \qquad \textit{with } d\check{\Pi}/d\Pi = \check{\pi}$$

is the Kullback–Leibler divergence from $\check{\Pi}$ *to* Π.

PROOF. Immediate from [57], Proposition 5.1. □

The proposition shows that, given the data $\mathcal{D}_n$ and a prior $\Pi = \Pi_n$ on g, the quasi-posterior $\Pi_n(dg \mid \mathcal{D}_n)$ defined in (4) is obtained as a minimizer of the empirical information complexity defined by (5) with $\eta = 1/2$. This gives a rational to use $\Pi_n(dg \mid \mathcal{D}_n)$ as a quasi-posterior since, among all possible "quasi-posteriors", this $\Pi_n(dg \mid \mathcal{D}_n)$ optimally balances the average of the natural loss function $g \mapsto \sum_{i=1}^{n} \hat{m}^2(W_i, g)$ and its complexity (or deviation) relative to the initial prior distribution measured by the Kullback–Leibler divergence. The scaling constant ("temperature") η is typically treated as a fixed constant (see, e.g., [35, 58]). An alternative way is to choose η in a data-dependent manner, by, for example, cross validation as mentioned in [58]. It is not difficult to see that the theory below can be extended to the case where η is even random, as long as η converges in probability to a fixed positive constant. However, for the sake of simplicity, we take $\eta = 1/2$ as a benchmark choice (note that as long as η is a fixed positive constant, the analysis can be reduced to the case with $\eta = 1/2$ by renormalization).

The quasi-posterior distribution provides point estimators of g_0. A most natural estimator would be the estimator defined by the posterior expectation (the expectation of the quasi-posterior distribution), that is,

$$\hat{g}_{\mathrm{QB}} = \begin{cases} \int g\, \Pi_n(dg \mid \mathcal{D}_n), & \text{if the right integral exists,} \\ 0, & \text{otherwise,} \end{cases} \tag{6}$$

where the integral $\int g\, \Pi_n(dg \mid \mathcal{D}_n)$ is understood as pointwise.

REMARK 1. Quasi-Bayesian approaches (not necessarily in the present form) are widely used and there are several other attempts of making probabilistic interpretation of such approaches. See, for example, [37] where the "limited information likelihood" is derived as the "best" (in a suitable sense) approximation to the true likelihood function under a set of moment restrictions and the Bayesian analysis with the limited information likelihood is argued ([44] adapted this approach to conditional moment restriction models), and [47] where a version of the empirical likelihood is interpreted in a Bayesian framework.

3. Main results. In this section, we study the asymptotic properties of the quasi-posterior distribution and the quasi-Bayes estimator. In doing so, we have to specify certain regularity properties, such as the smoothness of g_0 and the degree of ill-posedness of the problem. How to characterize the "smoothness" of g_0 is important since it is related to how to put priors. For this purpose, we find wavelet theory useful, and use sieve spaces constructed by using wavelet bases.

3.1. *Posterior construction.* To construct quasi-posterior distributions, we have to estimate $m(\cdot, g)$ and construct a sequence of sieve spaces for $\mathcal{G}$ on which priors concentrate. For the former purpose, we use a (wavelet) series estimator of $m(\cdot, g)$, as in [2] and [10]. For the latter purpose, we construct a sequence of sieve spaces formed by the wavelet basis.

We begin with stating the parameter space for g_0 and the wavelet basis used. We assume that the parameter space $\mathcal{G}$ is either $(B^s_{\infty,\infty}, \|\cdot\|_{s,\infty,\infty})$ (Hölder–Zygmund space) or $(B^s_{2,2}, \|\cdot\|_{s,2,2})$ (Sobolev space), where $B^s_{p,q}$ is the Besov space of functions on $[0, 1]$ with parameter (s, p, q) (the parameter s generally corresponds to "smoothness;" we add "s" on the parameter space, $\mathcal{G} = \mathcal{G}^s$, to clarify its dependence on s). See Appendix A.2 in the supplemental file [36] for the definition of Besov spaces. We assume that $s > 1/2$, under which $\mathcal{G}^s \subset C[0, 1]$.

Fix (sufficiently large) $J_0 \geq 0$, and let $\{\varphi^{\mathrm{int}}_{J_0 k}\}_{k=0}^{2^{J_0}-1} \cup \{\psi^{\mathrm{int}}_{jk}, j \geq J_0, k = 0, \ldots, 2^j - 1\}$ be an S-regular Cohen–Daubechies–Vial (CDV) wavelet basis for $L_2[0, 1]$ [14], where S is a positive integer larger than s. See Appendix A.1 in the supplemental file [36] for CDV wavelet bases. For the notational convenience, we write $\phi_1 = \varphi^{\mathrm{int}}_{J_0,0}, \phi_2 = \varphi^{\mathrm{int}}_{J_0,1}, \ldots, \phi_{2^{J_0}} = \varphi^{\mathrm{int}}_{J_0,2^{J_0}-1}$, and $\phi_{2^j+1} = \psi^{\mathrm{int}}_{j,0}, \phi_{2^j+2} = \psi^{\mathrm{int}}_{j,1}, \ldots, \phi_{2^{j+1}} = \psi^{\mathrm{int}}_{j,2^j-1}$ for $j \geq J_0$. Here and in what follows:

Take and fix an S-regular CDV wavelet basis of $\{\phi_l, l \geq 1\}$ with $S > s$,

and we keep this convention. Let V_j be the linear subspace of $L_2[0, 1]$ spanned by $\{\phi_1, \ldots, \phi_{2^j}\}$, and denote by P_j the projection operator onto V_j, that is, for any $g = \sum_{l=1}^{\infty} b_l \phi_l \in L_2[0, 1]$, $P_j g = \sum_{l=1}^{2^j} b_l \phi_l$. In what follows, for any $J \in \mathbb{N}$, the notation b^J means that it is a vector of dimension 2^J. For example, $b^J = (b_1, \ldots, b_{2^J})^T$.

REMARK 2 (Approximation property). For either $g \in B^s_{\infty,\infty}$ or $B^s_{2,2}$, we have $\|g - P_J g\|^2 \leq C 2^{-2Js}$ for all $J \geq J_0$. Here the constant C depends only on s and the corresponding Besov norm of g.

REMARK 3. The use of CDV wavelet bases is not crucial and one may use other reasonable bases such as the Fourier and Hermite polynomial bases. The theory below can be extended to such bases with some modifications. However, CDV wavelet bases are particularly well suited to approximate (not necessarily periodic) smooth functions, which is the reason why we use here CDV wavelet bases. On the other hand, for example, the Fourier basis is only appropriate to approximate periodic functions and it is often not natural to assume that the structural function g_0 is periodic.

We shall now move to the posterior construction. For $J \geq J_0$, define the 2^J-dimensional vector of functions $\phi^J(w)$ by

$$\phi^J(w) = \big(\phi_1(w), \ldots, \phi_{2^J}(w)\big)^T.$$

Let $J_n \geq J_0$ be a sequence of positive integers such that $J_n \to \infty$ and $2^{J_n} = o(n)$. Then a wavelet series estimator of $m(\cdot, g)$ is defined as

$$\hat{m}(w, g) = \phi^{J_n}(w)^T \big(\mathbb{E}_n[\phi^{J_n}(W_i)^{\otimes 2}]\big)^{-1} \mathbb{E}_n\big[\phi^{J_n}(W_i)\big(Y_i - g(X_i)\big)\big],$$

where we replace the inverse matrix by the generalized inverse if the former does not exist; the probability of such an event converges to zero as $n \to \infty$ under the assumptions below. We use this wavelet series estimator throughout the analysis.

For the same J_n, we shall take $V_{J_n} = \operatorname{span}\{\phi_1, \ldots, \phi_{2^{J_n}}\}$ as a sieve space for $\mathcal{G}^s$. We consider priors Π_n that concentrate on V_{J_n}, that is, $\Pi_n(V_{J_n}) = 1$. Formally, we think of that priors on g are defined on the Borel σ-field of $C[0, 1]$ (hence the quasi-posterior $\Pi_n(dg \mid \mathcal{D}_n)$ is understood to be defined on the Borel σ-field of $C[0, 1]$, which is possible since the map $g \mapsto p_g(\mathcal{D}_n)$ is continuous on $C[0, 1]$). Since the map $b^{J_n} = (b_1, \ldots, b_{2^{J_n}})^T \mapsto \sum_{l=1}^{2^{J_n}} b_l \phi_l, \mathbb{R}^{2^{J_n}} \to C[0, 1]$, is homeomorphic from $\mathbb{R}^{2^{J_n}}$ onto V_{J_n}, putting priors on $g \in V_{J_n}$ is equivalent to putting priors on $b^{J_n} \in \mathbb{R}^{2^{J_n}}$ (the latter are of course defined on the Borel σ-field of $\mathbb{R}^{2^{J_n}}$). Practically, priors on $g \in V_{J_n}$ are induced from priors on $b^{J_n} \in \mathbb{R}^{2^{J_n}}$. For the later purpose, it is useful to determine the correspondence between priors for these two parameterizations. Unless otherwise stated, we follow the convention of the notation such that

$$\tilde{\Pi}_n\text{: a prior on } b^{J_n} \in \mathbb{R}^{2^{J_n}} \leftrightarrow \Pi_n\text{: the induced prior on } g \in V_{J_n}.$$

We shall call $\tilde{\Pi}_n$ a generating prior, and Π_n the induced prior.

Correspondingly, the quasi-posterior for b^{J_n} is defined. With a slight abuse of notation, for $g = \sum_{l=1}^{2^{J_n}} b_l \phi_l$, we write $\hat{m}(w, b^{J_n}) = \hat{m}(w, g)$, and take $p_{b^{J_n}}(\mathcal{D}_n) = \exp\{-(n/2)\mathbb{E}_n[\hat{m}^2(W_i, b^{J_n})]\}$ as a quasi-likelihood for b^{J_n}. Note that in this particular setting, the log quasi-likelihood is quadratic in b^{J_n}. Let $\tilde{\Pi}_n(db^{J_n} \mid \mathcal{D}_n)$ denote the resulting quasi-posterior distribution for b^{J_n}:

$$\tilde{\Pi}_n\big(db^{J_n} \mid \mathcal{D}_n\big) = \frac{p_{b^{J_n}}(\mathcal{D}_n)\tilde{\Pi}_n(db^{J_n})}{\int p_{b^{J_n}}(\mathcal{D}_n)\tilde{\Pi}_n(db^{J_n})}. \tag{7}$$

For the quasi-Bayes estimator $\hat{g}_{\mathrm{QB}}$ defined by (6), since for every $x \in [0, 1]$, the map $g \mapsto g(x)$ is continuous on $C[0, 1]$, and conditional on $\mathcal{D}_n$ the quasi-posterior $\Pi_n(dg \mid \mathcal{D}_n)$ is a Borel probability measure on $C[0, 1]$, the integral

$\int g(x)\Pi_n(dg \mid \mathcal{D}_n)$ exists as soon as $\int |g(x)|\Pi_n(dg \mid \mathcal{D}_n) < \infty$. Furthermore, $\hat{g}_{\mathrm{QB}}$ can be computed by using the relation

$$\int g(x)\Pi_n(dg \mid \mathcal{D}_n) = \phi^{J_n}(x)^T\left[\int b^{J_n}\tilde{\Pi}_n(db^{J_n} \mid \mathcal{D}_n)\right]$$

as soon as the integral on the right-hand side exists. Hence, practically, it is sufficient to compute the expectation of $\tilde{\Pi}_n(db^{J_n} \mid \mathcal{D}_n)$.

REMARK 4. The use of the same wavelet basis to estimate $m(\cdot, g)$ and to construct a sequence of sieve spaces for $\mathcal{G}^s$ is not essential and can be relaxed. Suppose that we have another CDV wavelet basis $\{\tilde{\phi}_l\}$ for $L_2[0,1]$ and use this basis to estimate $m(\cdot, g)$. Then, all the results below apply by simply replacing $\phi_l(W_i)$ by $\tilde{\phi}_l(W_i)$. To keep the notation simple, we use the same wavelet basis.

However, the use of the same resolution level J_n is essential (at least at the proof level) in establishing the asymptotic properties of the quasi-posterior distribution. It may be a technical artifact, but we do not extend the theory in this direction since there is no clear theoretical benefit to do so (note that in the purely frequentist estimation case, [10] allowed for using different cut-off levels for approximating $m(\cdot, g)$ and $g(\cdot)$).

3.2. *Basic assumptions.* We state some basic assumptions. We do not state here assumptions on priors, which will be stated in the theorems below. In what follows, let $C_1 > 1$ be a sufficiently large constant.

ASSUMPTION 1. (i) (X, W) has a joint density $f_{X,W}(x, w)$ on $[0,1]^2$ satisfying that $f_{X,W}(x,w) \le C_1, \forall x, w \in [0,1]$. (ii) $\sup_{w\in[0,1]} \mathbb{E}[U^2 \mid W = w] \le C_1$ where $U = Y - g_0(X)$. (iii) $s_{\min}(\mathbb{E}[\phi^J(W)^{\otimes 2}]) \ge C_1^{-1}, \forall J \ge J_0$.

Assumption 1 is a usual restriction in the literature, up to minor differences (see [28, 32]). Denote by $f_X(x)$ and $f_W(w)$ the marginal densities of X and W, respectively, that is, $f_X(x) = \int f_{X,W}(x,w)\,dw$ and $f_W(w) = \int f_{X,W}(x,w)\,dx$. Then Assumption 1(i) implies that $f_X(x) \le C_1, \forall x \in [0,1]$ and $f_W(w) \le C_1, \forall w \in [0,1]$. A primitive regularity condition that guarantees Assumption 1(iii) is that $f_W(w) \ge C_1^{-1}$ for all $w \in [0,1]$. To see this, for $\alpha^J \in \mathbb{R}^{2^J}$ with $\|\alpha^J\|_{\ell^2} = 1$, we have

$$\begin{aligned}
(\alpha^J)^T\mathbb{E}[\phi^J(W)^{\otimes 2}]\alpha^J &= \int_0^1 (\phi^J(w)^T\alpha^J)^2 f_W(w)\,dw \\
&\ge C_1^{-1}\int_0^1 (\phi^J(w)^T\alpha^J)^2\,dw \\
&= C_1^{-1}(\alpha^J)^T\left[\int_0^1 \phi^J(w)\phi^J(w)^T\,dw\right]\alpha^J \\
&= C_1^{-1}\|\alpha^J\|_{\ell^2}^2 = C_1^{-1},
\end{aligned} \tag{8}$$

where we have used the fact that $\{\phi_l\}$ is orthonormal in $L_2[0,1]$.

For identification of g_0, we assume:

ASSUMPTION 2. The linear operator $K : L_2[0,1] \to L_2[0,1]$ is injective.

For smoothness of g_0, as mentioned before, we assume:

ASSUMPTION 3. $\exists s > 1/2$, $g_0 \in \mathcal{G}^s$, where $\mathcal{G}^s$ is either $B^s_{\infty,\infty}$ or $B^s_{2,2}$.

The identification condition (Assumption 2) is equivalent to the "completeness" of the conditional distribution of X conditional on W [45]. We refer the reader to [17, 49] and [34] for discussion on the completeness condition. We should note that restricting the domain of K to a "small" set, such as a Sobolev ball, would substantially relax Assumption 2, which however requires a different analysis. For the sake of simplicity, we assume the injectivity of K on the full domain.

As discussed in the Introduction, solving (2) is an ill-posed inverse problem. Thus, the statistical difficulty of estimating g_0 depends on the difficulty of continuously inverting K, which is usually referred to as "ill-posedness" of the inverse problem (2). Typically, the ill-posedness is characterized by the decay rate of $\kappa_l \to 0$ (κ_l is the lth largest singular value of K), which is plausible if K were known and the singular value decomposition of K were used (see [9]). However, here, K is unknown and the known wavelet basis $\{\phi_l\}$ is used instead of the singular value system. Thus, it is suitable to quantify the ill-posedness using the wavelet basis $\{\phi_l\}$. To this end, define

$$\tau_J = s_{\min}\big(\mathbb{E}[\phi^J(W)\phi^J(X)^T]\big) = s_{\min}\big((\langle \phi_l, K\phi_m\rangle)_{1\le l,m\le 2^J}\big), \qquad J \ge J_0.$$

This quantity corresponds to (the reciprocal of) what is called "sieve measure of ill-posedness" in the literature [5, 32]. We at least have to assume that $\tau_J > 0$ for all $J \ge J_0$. Note however that

$$\begin{aligned}
\tau_J &= s_{\min}\big((\langle \phi_l, K\phi_m\rangle)_{1\le l,m\le 2^J}\big) \\
&= \min_{g\in V_J, \|g\|=1} \big\|(\langle \phi_l, Kg\rangle)_{1\le l\le 2^J}\big\|_{\ell^2} \\
&\le \min_{g\in V_J, \|g\|=1} \|Kg\| \qquad \text{(Bessel's inequality)} \\
&\le \kappa_{2^J} \qquad \text{(Courant–Fischer–Weyl's minimax principle)}
\end{aligned}$$

by which, necessarily, $\tau_J \to 0$ as $J \to \infty$. For this quantity, we assume:

ASSUMPTION 4. (i) (Mildly ill-posed case) $\exists r > 0$, $\tau_J \ge C_1^{-1} 2^{-Jr}$, $\forall J \ge J_0$ or (severely ill-posed case) $\exists c > 0$, $\tau_J \ge C_1^{-1}\exp(-c2^J)$, $\forall J \ge J_0$;

(ii)

$$\begin{aligned}
&\|\mathbb{E}[\phi^J(W)(g_0 - P_J g_0)(X)]\|_{\ell^2} \big(= \|(\langle \phi_l, K(g_0 - P_J g_0)\rangle)_{l=1}^{2^J}\|_{\ell^2}\big) \\
&\quad \le C_1 \tau_J \|g_0 - P_J g_0\| \qquad \forall J \ge J_0.
\end{aligned}$$

Assumption 4(i) lower bounds τ_J as $J \to \infty$, thereby quantifies the ill-posedness. We cover both the "mildly ill-posed" and "severely ill-posed" cases (this definition of mild ill-posedness and severe ill-posedness is due to [31, 32]). The severely ill-posed case happens, for example, when the joint density $f_{X,W}(x, w)$ is analytic (see [43], Theorem 15.20).

Assumption 4(ii) is a "stability" condition about the bias $g_0 - P_J g_0$, which states that $K(g_0 - P_J g_0)$ is sufficiently "small" relative to $g_0 - P_J g_0$. Note that in the (ideal) case in which, for example, K is self-adjoint and $\{\phi_l\}$ is the eigen-basis of K, $\langle \phi_l, K(g_0 - P_J g_0)\rangle = 0$ for all $l = 1, \dots, 2^J$, in which case Assumption 4(ii) is trivially satisfied. Assumption 4(ii) allows more general situations in which K may not be self-adjoint and $\{\phi_l\}$ may not be the eigen-basis of K by allowing for a certain "slack." This assumption, although looks technical, is common in the study of rates of convergence in estimation of the structural function g_0. Indeed, essentially similar conditions have appeared in the past literature such as [5, 11, 32]. For example, [5], Assumption 6, essentially states (in our notation) that $\|K(g_0 - P_J g_0)\| \le C_1 \tau_J \|g_0 - P_J g_0\|$, which implies our Assumption 4(ii) since $\|(\langle \phi_l, K(g_0 - P_J g_0)\rangle)_{l=1}^{2^J}\|_{\ell^2} \le \|K(g_0 - P_J g_0)\|$ (Bessel's inequality).

REMARK 5. For given values of $C_1 > 1, M > 0, r > 0, c > 0$ and $s > 1/2$, let $\mathcal{F} = \mathcal{F}(C_1, M, r, c, s)$ denote the set of all distributions of (Y, X, W) satisfying Assumptions 1–4 with $\|g_0\|_{s,\infty,\infty} \le M$ in case of $\mathcal{G}^s = B^s_{\infty,\infty}$ and $\|g_0\|_{s,2,2} \le M$ in case of $\mathcal{G}^s = B^s_{2,2}$. By [11, 28], it is shown that the minimax rate of convergence (in $\|\cdot\|$) of estimation of g_0 over this distribution class $\mathcal{F}$ is $n^{-s/(2r+2s+1)}$ in the mildly ill-posed case (where $\tau_J \ge C_1^{-1} 2^{-Jr}$) and $(\log n)^{-s}$ in the severely ill-posed case [where $\tau_J \ge C_1^{-1} \exp(-c 2^J)$] as the sample size $n \to \infty$ (the assumption on the conditional second moment of U given W is not binding; that is, replacing Assumption 1(ii) by a stronger one, such as $\sup_{w \in [0,1]} \mathbb{E}[|U|^{2+\epsilon} \mid W = w] \le C_1$ for some $\epsilon > 0$ determined outside the class of distributions, does not alter these minimax rates).

By Theorem 2.5 of [24], it is readily seen that these rates are the fastest possible rates of contraction of (general) quasi-posterior distributions in this setting. More formally, we can state the following assertion:

Let $\Pi_n(dg \mid \mathcal{D}_n)$ be the quasi-posterior distribution defined on, say, the Borel σ-field of $C[0, 1]$, constructed from putting a suitable prior on g to the quasi-likelihood $p_g(\mathcal{D}_n)$ (the prior here needs not be a sieve prior). Suppose now that for some $\varepsilon_n \to 0$, $\sup_{F \in \mathcal{F}} \mathbb{E}_F[\Pi_n(g : \|g - g_0\| > \varepsilon_n \mid \mathcal{D}_n)] \to 0$. Then there exists a point estimator that converges (in probability) at least as fast as ε_n uniformly in $F \in \mathcal{F}$.

The proof is just a small modification of that of Theorem 2.5 in [24] and hence omitted. Importantly, the quasi-posterior cannot contract at a rate faster than the optimal rate of convergence for point estimators ([24], page 507, lines 19–20).

Hence, in the minimax sense, the fastest possible rate of contraction of the quasi-posterior distribution $\Pi_n(dg \mid \mathcal{D}_n)$ is $n^{-s/(2r+2s+1)}$ in the mildly ill-posed case and $(\log n)^{-s}$ in the severely ill-posed case (Proposition 2 in Section 4 ahead shows that these rates are indeed attainable for suitable sieve priors).

3.3. *Main results*: *General theorems*. This section presents general theorems on contraction rates and asymptotic normality for the quasi-posterior distribution as well as convergence rates for the quasi-Bayes estimator. In what follows, let $(Y_1, X_1, W_1), \dots, (Y_n, X_n, W_n)$ be i.i.d. observations of (Y, X, W). Denote by $b_0^J = (b_{01}, \dots, b_{0,2^J})^T$ the vector of the first 2^J generalized Fourier coefficients of g_0, that is, $b_{0l} = \int \phi_l g_0$. Let $\| \cdot \|_{\mathrm{TV}}$ denote the total variation norm between two distributions.

THEOREM 1. *Suppose that Assumptions 1–4 are satisfied. Take J_n in such a way that $J_n \to \infty$ and $J_n 2^{J_n}/n = o(\tau_{J_n}^2)$. Let ϵ_n be a sequence of positive constants such that $\epsilon_n \to 0$ and $n\epsilon_n^2 \gtrsim 2^{J_n}$. Suppose that generating priors $\tilde{\Pi}_n$ has densities $\tilde{\pi}_n$ on $\mathbb{R}^{2^{J_n}}$ and satisfy the following conditions*:

(P1) (*Small ball condition*). *There exists a constant $C > 0$ such that for all n sufficiently large, $\tilde{\Pi}_n(b^{J_n} : \|b^{J_n} - b_0^{J_n}\|_{\ell^2} \le \epsilon_n) \ge e^{-Cn\epsilon_n^2}$.*

(P2) (*Prior flatness condition*). *Let $\gamma_n = 2^{-J_n s} + \tau_{J_n}^{-1}\epsilon_n$. There exists a sequence of constants $L_n \to \infty$ sufficiently slowly such that for all n sufficiently large, $\tilde{\pi}_n(b^{J_n})$ is positive for all $\|b^{J_n} - b_0^{J_n}\|_{\ell^2} \le L_n\gamma_n$, and*

$$\sup_{\|b^{J_n}\|_{\ell^2} \le L_n\gamma_n, \|\tilde{b}^{J_n}\|_{\ell^2} \le L_n\gamma_n} \left| \frac{\tilde{\pi}_n(b_0^{J_n} + b^{J_n})}{\tilde{\pi}_n(b_0^{J_n} + \tilde{b}^{J_n})} - 1 \right| \to 0.$$

Then for every sequence $M_n \to \infty$, we have

$$\tilde{\Pi}_n\{b^{J_n} : \|b^{J_n} - b_0^{J_n}\|_{\ell^2} > M_n(2^{-J_n s} + \tau_{J_n}^{-1}\sqrt{2^{J_n}/n}) \mid \mathcal{D}_n\} \xrightarrow{P} 0. \tag{9}$$

Furthermore, assume that $J_n 2^{3J_n}/n = o(\tau_{J_n}^2)$. Then we have

$$\|\tilde{\Pi}_n(\cdot \mid \mathcal{D}_n) - N(\hat{b}^{J_n}, n^{-1}\Phi_{WX}^{-1}\Phi_{WW}\Phi_{XW}^{-1})(\cdot)\|_{\mathrm{TV}} \xrightarrow{P} 0,$$

where $\Phi_{WX} := \mathbb{E}[\phi^{J_n}(W)\phi^{J_n}(X)^T]$, $\Phi_{XW} := \Phi_{WX}^T$, $\Phi_{WW} := \mathbb{E}[\phi^{J_n}(W)^{\otimes 2}]$, and where $\hat{b}^{J_n}$ is a "maximum quasi-likelihood estimator" of $b_0^{J_n}$, that is,

$$\hat{b}^{J_n} \in \arg \max_{b^{J_n} \in \mathbb{R}^{2^{J_n}}} p_{b^{J_n}}(\mathcal{D}_n). \tag{10}$$

PROOF. See Section 5.1. $\square$

REMARK 6. The condition $J_n 2^{J_n}/n = o(\tau_{J_n}^2)$ appears essentially because the operator K is unknown. In our setup, this results in estimating the matrix $\mathbb{E}[\phi^{J_n}(W)\phi^{J_n}(X)^T]$ by its empirical counterpart $\mathbb{E}_n[\phi^{J_n}(W_i)\phi^{J_n}(X_i)^T]$. In the proof, we have to suitably lower bound the minimum singular value of $\mathbb{E}_n[\phi^{J_n}(W_i)\phi^{J_n}(X_i)^T]$, denoted by $\hat{\tau}_{J_n}$, which is an empirical counterpart of the sieve measure of ill-posedness τ_{J_n}. By Lemma 1, we have $\hat{\tau}_{J_n} = \tau_{J_n} - O_P(\sqrt{J_n 2^{J_n}/n})$, so that to make the estimation effect in $\hat{\tau}_{J_n}$ negligible, we need $J_n 2^{J_n}/n = o(\tau_{J_n}^2)$.

REMARK 7. Theorem 1 is abstract in the sense that it only gives conditions (P1) and (P2) on priors for which (9) and (10) hold. For specific priors, we have to check these conditions with possible J_n, which will be done in Section 4.

Since for $g = \sum_{l=1}^{2^{J_n}} b_l \phi_l$, $\|g - g_0\|^2 = \|g - P_{J_n} g_0\|^2 + \|g_0 - P_{J_n} g_0\|^2 \lesssim \|b^{J_n} - b_0^{J_n}\|_{\ell^2}^2 + 2^{-2J_n s}$, part (9) of Theorem 1 leads to that for every sequence $M_n \to \infty$, we have

$$\Pi_n\{g : \|g - g_0\| > M_n(2^{-J_n s} + \tau_{J_n}^{-1}\sqrt{2^{J_n}/n}) \mid \mathcal{D}_n\} \xrightarrow{P} 0,$$

which means that the rate of contraction of the quasi-posterior distribution $\Pi_n(dg \mid \mathcal{D}_n)$ is $\max\{2^{-J_n s}, \tau_{J_n}^{-1}\sqrt{2^{J_n}/n}\}$.[6] In many examples, for given $J_n \to \infty$ with $J_n 2^{J_n}/n = o(\tau_{J_n}^2)$, condition (P1) is satisfied with $\epsilon_n \sim \sqrt{2^{J_n}(\log n)/n}$. Taking J_n in such a way that [with some constant $c' < 1/(2c)$ in the severely ill-posed case]

$$\begin{cases} 2^{J_n} \sim n^{1/(2r+2s+1)}, & \text{in the mildly ill-posed case,} \\ \lim_{n\to\infty} (2^{J_n}/(c' \log n)) = 1, & \text{in the severely ill-posed case,} \end{cases} \tag{11}$$

under which the optimal contraction rate is attained, γ_n in condition (P2) is

$$\gamma_n \sim \begin{cases} n^{-s/(2r+2s+1)}(\log n)^{1/2}, & \text{in the mildly ill-posed case,} \\ (\log n)^{-s}, & \text{in the severely ill-posed case.} \end{cases} \tag{12}$$

So condition (P2) states that, to attain the optimal contraction rate (and the Bernstein–von Mises type result), the prior density $\tilde{\pi}_n$ should be sufficiently "flat" in a ball with center $b_0^{J_n}$ and radius of order (12). Some specific priors leading to the optimal contraction rate will be given in Section 4.

As noted before, in many examples, for given $J_n \to \infty$ with $J_n 2^{J_n}/n = o(\tau_{J_n}^2)$, condition (P1) is satisfied with $\epsilon_n \sim \sqrt{2^{J_n}(\log n)/n}$. Inspection of the proof shows that, without condition (P2), this already leads to contraction rate $\max\{2^{-J_n s}, \tau_{J_n}^{-1}\sqrt{2^{J_n}(\log n)/n}\}$, which, in the mildly ill-posed case, reduces to

[6] We have ignored the appearance of $M_n \to \infty$, which can be arbitrarily slow. A version in which M_n is replaced by a large fixed constant $M > 0$ is presented in Theorem 2.

$(n/\log n)^{-s/(2r+2s+1)}$ by taking $2^{J_n} \sim (n/\log n)^{1/(2r+2s+1)}$. However, this rate is not fully satisfactory because of the appearance of the log term. Condition (P2) is used to get rid of the log term.

The small ball condition (P1) is standard in nonparametric Bayesian statistics and analogous to condition (2.4) in [24]. It is, however, stated in [24], pages 505–506, that their Theorem 2.1 is not sharp enough when priors constructed on a sequence of finite-dimensional sieves are used, and the more sophisticated condition (2.9) is devised in their Theorem 2.4 (see also the proof of their Theorem 4.5). However, a version of their condition (2.9) is not clear to work in our problem, because the effect of the random matrix $\mathbb{E}_n[\phi^{J_n}(W_i)\phi^{J_n}(X_i)^T]$ has to be suitably controlled. Instead, we devise condition (P2) to obtain sharper contraction rates.

Under a further integrability condition about $U = Y - g_0(X)$, $M_n \to \infty$ in (9) can be replaced by a large fixed constant M.

THEOREM 2. *Suppose that all the conditions that guarantee* (9) *in Theorem* 1 *are satisfied. Furthermore, assume that* $\sup_{w\in[0,1]} \mathbb{E}[U^2 1(|U| > \lambda) \mid W = w] \to 0$ *as* $\lambda \to \infty$ *where* $U = Y - g_0(X)$. *Then there exists a constant* $M > 0$ *such that*

$$\tilde{\Pi}_n\{b^{J_n} : \|b^{J_n} - b_0^{J_n}\|_{\ell^2} > M(2^{-J_n s} + \tau_{J_n}^{-1}\sqrt{2^{J_n}/n}) \mid \mathcal{D}_n\} \xrightarrow{P} 0. \tag{13}$$

PROOF. See Section 5.2. $\square$

The proof consists in establishing a concentration property of the random variable $\|\mathbb{E}_n[\phi^{J_n}(W_i)U_i]\|_{\ell^2}$, which uses a truncation argument and Talagrand's [52] concentration inequality. A sufficient condition that guarantees that

$$\sup_{w\in[0,1]} \mathbb{E}[U^2 1(|U| > \lambda) \mid W = w] \to 0$$

as $\lambda \to \infty$ is that $\exists\epsilon > 0$, $\sup_{w\in[0,1]} \mathbb{E}[|U|^{2+\epsilon} \mid W = w] < \infty$. The additional condition in Theorem 2 is a uniform integrability condition and stronger than Assumption 1(ii). To see this, note that U is distributed as $F_{U|W}^{-1}(\mathcal{U} \mid W)$ where $F_{U|W}^{-1}(u \mid w)$ is the conditional quantile function of U given $W = w$, and $\mathcal{U}$ is a uniform random variable on (0, 1) independent of W. Think of $U_w(u) = F_{U|W}^{-1}(u \mid w)$, $w \in [0, 1]$ as a stochastic process defined on the probability space $((0, 1), \mu)$ with μ Lebesgue measure on (0, 1). Then the condition $\sup_{w\in[0,1]} \mathbb{E}[U^2 1(|U| > \lambda) \mid W = w] \to 0$ as $\lambda \to \infty$ states exactly the uniform integrability of $(U_w)_{w\in[0,1]}$.

The second part of Theorem 1 states a Bernstein–von Mises type result for the quasi-posterior distribution $\tilde{\Pi}_n(db^{J_n} \mid \mathcal{D}_n)$, which states that the quasi-posterior distribution is approximated by the normal distribution centered at $\hat{b}^{J_n}$, which is often referred to as the "sieve minimum distance estimator" and is a benchmark frequentist estimator for these types of models. Note that, neglecting the bias, $\hat{b}^{J_n}$ is approximated as $b_0^{J_n} + \Phi_{WX}^{-1}\mathbb{E}_n[\phi^{J_n}(W_i)U_i]$, but the covariance matrix of the term $\Phi_{WX}^{-1}\sqrt{n}\mathbb{E}_n[\phi^{J_n}(W_i)U_i]$ is generally different from $\Phi_{WX}^{-1}\Phi_{WW}\Phi_{XW}^{-1}$ (which is the reason why we added "type"). This is a generic nature of quasi-posterior distributions. Even for finite-dimensional models, generally, the covariance matrix of the centering variable does not coincide with that of the normal distribution approximating the quasi-posterior distribution (see [12]).

Finally, we consider the convergence rate of the quasi-Bayes estimator $\hat{g}_{\mathrm{QB}}$ of g_0 defined by (6).

THEOREM 3. *Suppose that all the conditions of Theorem 2 are satisfied. Let $\hat{g}_{\mathrm{QB}}$ be the quasi-Bayes estimator defined by* (6). *Then* $\mathbb{P}\{\mathcal{D}_n : \int |g(x)|\Pi_n(dg \mid \mathcal{D}_n) < \infty, \forall x \in [0,1]\} \to 1$, *and there exists a constant* $D > 0$ *such that for every sequence* $M_n \to \infty$,

$$\mathbb{P}\big[\|\hat{g}_{\mathrm{QB}} - g_0\| \le D \max\{2^{-J_n s}, \tau_{J_n}^{-1}\sqrt{2^{J_n}/n}, \tau_{J_n}^{-1}\epsilon_n \varrho_n M_n\}\big] \to 1, \tag{14}$$

where

$$\varrho_n := \sup_{\|b^{J_n}\|_{\ell^2} \le L_n\gamma_n, \|\tilde{b}^{J_n}\|_{\ell^2} \le L_n\gamma_n} \left| \frac{\tilde{\pi}_n(b_0^{J_n} + b^{J_n})}{\tilde{\pi}_n(b_0^{J_n} + \tilde{b}^{J_n})} - 1 \right|,$$

and where ϵ_n, γ_n and L_n are given in the statement of Theorem 1.

PROOF. See Appendix C in the supplemental file [36]. □

Theorem 3 is *not* directly deduced from Theorem 1. Indeed, $\|g - g_0\|$ may be unbounded on the support of Π_n since the support of Π_n may be unbounded in $\|\cdot\|$, and hence the argument in [24], pages 506–507, cannot apply (in [24], a typical distance to measure the goodness of a point estimator is the Hellinger distance and uniformly bounded). Hence, additional work is needed to prove Theorem 3.

The convergence rate of the quasi-Bayes estimator is determined by the three terms: $2^{-J_n s}$, $\tau_{J_n}^{-1}\sqrt{2^{J_n}/n}$, and $\tau_{J_n}^{-1}\epsilon_n\varrho_n M_n$. The last term is typically small relative to the other two terms. Indeed, as noted before, in many examples, for given $J_n \to \infty$ with $J_n 2^{J_n}/n = o(\tau_{J_n}^2)$, ϵ_n can be taken in such a way that $\epsilon_n \sim \sqrt{2^{J_n}(\log n)/n}$. In that case $\tau_{J_n}^{-1}\epsilon_n\varrho_n M_n \sim \tau_{J_n}^{-1}\varrho_n M_n\sqrt{2^{J_n}(\log n)/n}$, and as long as $\varrho_n \to 0$ sufficiently fast, that is, $\varrho_n = o((\log n)^{-1/2})$, the convergence rate of the quasi-Bayes estimator $\hat{g}_{\mathrm{QB}}$ reduces to $\max\{2^{-J_n s}, \tau_{J_n}^{-1}\sqrt{2^{J_n}/n}\}$.

4. Prior specification: Examples. In this section, we give some specific sieve priors for which the quasi-posterior distribution (the quasi-Bayes estimator) attains the minimax optimal rate of contraction (convergence, resp.). We consider two types of priors, namely, product and isotropic priors. We will verify that these priors meet conditions (P1) and (P2) in Theorem 1 with the choice (11). For the notational convenience, define

$$\varepsilon_{n,s,r} = \begin{cases} n^{-s/(2s+2r+1)}, & \text{in the mildly ill-posed case,} \\ (\log n)^{-s}, & \text{in the severely ill-posed case.} \end{cases}$$

We may think of the severely ill-posed case as the case with $r = \infty$.

PROPOSITION 2. *Suppose that Assumptions 1–4 are satisfied. Consider the following two classes of prior distributions on* $\mathbb{R}^{2^{J_n}}$:

(Product prior). *Let* $q(x)$ *be a probability density function on* $\mathbb{R}$ *such that for a constant* $A > \sup_{l \ge 1} |b_{0l}|$: (1) $q(x)$ *is positive on* $[-A, A]$; (2) $\log q(x)$ *is Lipschitz continuous on* $[-A, A]$, *that is, there exists a constant* $L > 0$ *possibly depending on* A *such that* $|\log q(x) - \log q(y)| \le L|x - y|, \forall x, y \in [-A, A]$. *Take the density of the generating prior by* $\tilde{\pi}_n(b^{J_n}) = \prod_{l=1}^{2^{J_n}} q(b_l)$.

(Isotropic prior). *Let* $r(x)$ *be a probability density function on* $[0, \infty)$ *having all moments such that*: (1) *for a constant* $A > \|g_0\|$, $r(x)$ *is positive and continuous on* $[0, A]$; (2) *for a constant* $c'' > 0$, $\int_0^\infty x^{k-1} r(x)\,dx \le e^{c'' k \log k}$ *for all* k *sufficiently large. Take the density of the generating prior by* $\tilde{\pi}_n(b^{J_n}) \propto r(\|b^{J_n}\|_{\ell^2})$.

Take J_n *as in* (11). *Then, in either case of product or isotropic priors, for every sequence* $M_n \to \infty$, *we have* $\Pi_n\{g : \|g - g_0\| > M_n \varepsilon_{n,s,r} \mid \mathcal{D}_n\} \xrightarrow{P} 0$. *Furthermore, if* $\sup_{w \in [0,1]} \mathbb{E}[U^2 1(|U| > \lambda) \mid W = w] \to 0$ *as* $\lambda \to \infty$, *then there exists a constant* $M > 0$ *such that* $\Pi_n\{g : \|g - g_0\| > M \varepsilon_{n,s,r} \mid \mathcal{D}_n\} \xrightarrow{P} 0$.

PROOF. See Appendix D in the supplemental file [36]. □

Proposition 2 shows that a wide class of priors constructed on slowly growing sieves lead to the minimax optimal contraction rate (see Remark 5). In either case of product or isotropic priors, the constant A is not necessarily known, which allows $q(x)$ and $r(x)$ to have unbounded support. For example, in the former case, $q(x)$ may be the density of the standard normal distribution, in which case A can be taken to be arbitrarily large. Likewise, in the latter case, $r(x)$ may be the density of an exponential distribution: $r(x) = \lambda e^{-\lambda x}, x \ge 0$ for some $\lambda > 0$. In the isotropic prior case, $r(x)$ should have all moments, that is, $\int_0^\infty x^k r(x)\,dx < \infty$ for all $k \ge 1$, which ensures that $\tilde{\pi}_n(b^{J_n}) \propto r(\|b^{J_n}\|_{\ell^2})$ is a proper distribution on $\mathbb{R}^{2^{J_n}}$ for every $n \ge 1$.

The next proposition shows that two classes of priors in Proposition 2 lead to the minimax optimal convergence rate for the quasi-Bayes estimator.

PROPOSITION 3. *Suppose that Assumptions 1–4 are satisfied. Furthermore, assume that* $\sup_{w \in [0,1]} \mathbb{E}[U^2 1(|U| > \lambda) \mid W = w] \to 0$ *as* $\lambda \to \infty$. *Consider the two classes of prior distributions on* $\mathbb{R}^{2^{J_n}}$ *given in Proposition 2. In the isotropic prior case, assume further that* $r(x)$ *is Lipschitz continuous on* $[0, A]$. *Take* J_n *as in* (11). *Then, in either case of product or isotropic priors, there exists a constant* $M > 0$ *such that* $\mathbb{P}\{\|\hat{g}_{\mathrm{QB}} - g_0\| > M \varepsilon_{n,s,r}\} \to 0$.

PROOF. See Appendix D in the supplemental file [36]. □

REMARK 8. In the above propositions, J_n plays the role of regularization and should be chosen sufficiently slowly growing, thereby there is no need to place restrictions on weights on b_l between $1 \le l \le 2^{J_n}$. The abstract Theorem 1 is derived to cover this case. There is another way to deal with the ill-posedness, that is, allowing for large-dimensional sieves but placing prior distributions that have smaller weights on b_l for larger l ("shrinking priors"), which corresponds to the "sieve method using large-dimensional sieves with heavy penalties" in the classification of [10].[7] The supplementary material of [44] is concerned with this approach, but they did not establish sharp contraction rates. The extension to this approach requires a different technique than that used in the present paper, and remains as an open problem.

5. Proofs of Theorems 1 and 2.

5.1. *Proof of Theorem* 1. Before proving Theorem 1, we first prepare some technical lemmas (Lemmas 1–3) and establish preliminary rates of contraction for the quasi-posterior distribution (Proposition 4). Proofs of Lemmas 1–3 are given in Appendix B in the supplemental file [36]. For the notational convenience, define the matrices

$$\hat{\Phi}_{WX} = \mathbb{E}_n\big[\phi^{J_n}(W_i)\phi^{J_n}(X_i)^T\big], \qquad \hat{\Phi}_{XW} = \hat{\Phi}_{WX}^T,$$
$$\hat{\Phi}_{WW} = \mathbb{E}_n\big[\phi^{J_n}(W_i)^{\otimes 2}\big],$$

which are the empirical counterparts of Φ^{WX}, Φ^{XW} and Φ_{WW}, respectively. Also define

$$U_i = Y_i - g_0(X_i), \qquad R_i = Y_i - P_{J_n} g_0(X_i), \qquad \Delta_n = \sqrt{n}\mathbb{E}_n\big[\phi^{J_n}(W_i)R_i\big].$$

Lemma 1 is a technical lemma on these quantities. Lemma 2 characterizes the total variation convergence between two centered multivariate normal distributions with increasing dimensions in terms of the speed of convergence between the corresponding covariance matrices. Lemma 3 will be used in the latter part in the proof of Theorem 1.

LEMMA 1. *Suppose that Assumptions* 1–4 *are satisfied. Let* $J_n \to \infty$ *as* $n \to \infty$. (i) *There exists a constant* $D > 0$ *such that* $\sup_{w \in [0,1]} \|\phi^J(w)\|_{\ell^2} \le D 2^{J/2}$ *for all* $J \ge J_0$. (ii) $C_1^{-1} \le s_{\min}(\mathbb{E}[\phi^J(W)^{\otimes 2}]) \le s_{\max}(\mathbb{E}[\phi^J(W)^{\otimes 2}]) \le C_1$ *and* $s_{\max}(\mathbb{E}[\phi^J(W)\phi^J(X)^T]) \le C_1$ *for all* $J \ge J_0$. (iii) *If* $J_n 2^{J_n}/n \to 0$, $\|\hat{\Phi}_{WW} - \Phi_{WW}\|_{\mathrm{op}} = O_P(\sqrt{J_n 2^{J_n}/n})$ *and* $\|\hat{\Phi}_{WX} - \Phi_{WX}\|_{\mathrm{op}} = O_P(\sqrt{J_n 2^{J_n}/n})$.

[7] The previous version of this paper contains results on shrinking priors, but J_n should be still slowly growing as in the above propositions, which corresponds to the sieve method using slowly growing sieves with *light* penalties. Those results have been removed in the current version according to the referee's suggestion, but available upon request.

(iv) $\|\mathbb{E}_n[\phi^{J_n}(W_i)R_i]\|_{\ell^2}^2 = O_P(2^{J_n}/n + \tau_{J_n}^2 2^{-2J_n s})$. (v) *If* $J_n 2^{J_n}/n = o(\tau_{J_n}^2)$, $s_{\min}(\hat{\Phi}_{WX}) \geq (1 - o_P(1))\tau_{J_n}$.

LEMMA 2. *Let Σ_n be a sequence of symmetric positive definite matrices of dimension $k_n \to \infty$ as $n \to \infty$ such that $\|\Sigma_n - I_{k_n}\|_{\mathrm{op}} = o(k_n^{-1})$. Then as $n \to \infty$,*

$$\int \left| dN(0, \Sigma_n)(x) - dN(0, I_{k_n})(x) \right| dx \to 0.$$

LEMMA 3. *Let $\hat{A}_n$ be a sequence of random $k_n \times k_n$ matrices where k_n is either bounded or $k_n \to \infty$ as $n \to \infty$. Suppose that there exist sequences of positive constants ϵ_n, δ_n and a sequence of nonrandom, nonsingular $k_n \times k_n$ matrices A_n such that $\epsilon_n \to 0, \delta_n \to 0, s_{\min}(A_n) \gtrsim \epsilon_n, \|\hat{A}_n - A_n\|_{\mathrm{op}} = O_P(\delta_n)$ and $\epsilon_n^{-1}\delta_n \to 0$. Then $\hat{A}_n$ is nonsingular with probability approaching one and $\|\hat{A}_n^{-1}A_n - I_{k_n}\|_{\mathrm{op}} \vee \|A_n\hat{A}_n^{-1} - I_{k_n}\|_{\mathrm{op}} = O_P(\epsilon_n^{-1}\delta_n)$.*

The following proposition gives preliminary rates of contraction for the quasi-posterior distribution.

PROPOSITION 4 (Preliminary contraction rates). *Suppose that Assumptions 1–4 are satisfied. Take J_n in such a way that $J_n \to \infty$ and $J_n 2^{J_n}/n = o(\tau_{J_n}^2)$. Let ϵ_n be a sequence of positive constants such that $\epsilon_n \to 0$ and $\sqrt{n}\epsilon_n \to \infty$. Assume that a sequence of generating priors $\tilde{\Pi}_n$ satisfies condition* (P1) *of Theorem* 1. *Define the data-dependent, empirical seminorm $\|\cdot\|_{\mathcal{D}_n}$ on $\mathbb{R}^{2^{J_n}}$ by*

$$\|b^{J_n}\|_{\mathcal{D}_n} = \|\hat{\Phi}_{WX}b^{J_n}\|_{\ell^2}, \qquad b^{J_n} \in \mathbb{R}^{2^{J_n}}.$$

Then for every sequence $M_n \to \infty$, we have

$$\tilde{\Pi}_n\{b^{J_n} : \|b^{J_n} - b_0^{J_n}\|_{\mathcal{D}_n} > M_n(\epsilon_n + \tau_{J_n}2^{-J_n s}) \mid \mathcal{D}_n\} \xrightarrow{P} 0.$$

PROOF. Proof of Proposition 4 The proof consists of constructing suitable "tests" and is essentially similar to, for example, the proof of Theorem 2.1 in [24]. Let $\delta_n = \epsilon_n + \tau_{J_n}2^{-J_n s}$. We wish to show that there exists a constant $c_0 > 0$ such that

$$\mathbb{P}\{\tilde{\Pi}_n(b^{J_n} : \|b^{J_n} - b_0^{J_n}\|_{\mathcal{D}_n} > M_n\delta_n \mid \mathcal{D}_n) \leq e^{-c_0 M_n^2 n\delta_n^2}\} \to 1. \tag{15}$$

Note that since $\sqrt{n}\epsilon_n \to \infty$, $n\delta_n^2 \geq n\epsilon_n^2 \to \infty$. Below, $c_1, c_2, \ldots$ are some positive constants of which the values are understood in the context.

Note that $Y_i = P_{J_n}g_0(X_i) + R_i = \phi^{J_n}(X_i)^T b_0^{J_n} + R_i$. Then for $b^{J_n} \in \mathbb{R}^{2^{J_n}}$,

$$\begin{aligned}
\mathbb{E}_n[\hat{m}^2(W_i, b^{J_n})] = {} & -2(b^{J_n} - b_0^{J_n})^T \hat{\Phi}_{XW}\hat{\Phi}_{WW}^{-1}\mathbb{E}_n[\phi^{J_n}(W_i)R_i] \\
& + (b^{J_n} - b_0^{J_n})^T \hat{\Phi}_{XW}\hat{\Phi}_{WW}^{-1}\hat{\Phi}_{WX}(b^{J_n} - b_0^{J_n}) \\
& + \mathbb{E}_n[\phi^{J_n}(W_i)R_i]^T \hat{\Phi}_{WW}^{-1}\mathbb{E}_n[\phi^{J_n}(W_i)R_i].
\end{aligned} \tag{16}$$

Since the last term is independent of b^{J_n}, it is canceled out in the quasi-posterior distribution. Denote by $\ell_{b^{J_n}}(\mathcal{D}_n)$ the sum of the first two terms in (16). Then

$$\tilde{\Pi}_n(db^{J_n} \mid \mathcal{D}_n) \propto \exp\{-(n/2)\ell_{b^{J_n}}(\mathcal{D}_n)\}\tilde{\Pi}_n(db^{J_n}).$$

Using the fact that for any $x, y, c \in \mathbb{R}$ with $c > 0$, $2xy \le cx^2 + c^{-1}y^2$, we have

$$\begin{aligned}\ell_{b^{J_n}}(\mathcal{D}_n) &\ge (\hat{\lambda}_{\min} - c)\|b^{J_n} - b_0^{J_n}\|_{\mathcal{D}_n}^2 \\ &\quad - c^{-1}\hat{\lambda}_{\max}^2\|\mathbb{E}_n[\phi^{J_n}(W_i)R_i]\|_{\ell^2}^2 \qquad \forall c > 0,\end{aligned} \tag{17}$$

where $\hat{\lambda}_{\min}$ and $\hat{\lambda}_{\max}$ are the minimum and maximum eigenvalues of the matrix $\hat{\Phi}_{WW}^{-1}$, respectively. Likewise, we have

$$\begin{aligned}\ell_{b^{J_n}}(\mathcal{D}_n) &\le (\hat{\lambda}_{\max} + c)\|b^{J_n} - b_0^{J_n}\|_{\mathcal{D}_n}^2 \\ &\quad + c^{-1}\hat{\lambda}_{\max}^2\|\mathbb{E}_n[\phi^{J_n}(W_i)R_i]\|_{\ell^2}^2 \qquad \forall c > 0.\end{aligned} \tag{18}$$

Define the event

$$\begin{aligned}\mathcal{E}_{1n} = &\{\mathcal{D}_n : \hat{\lambda}_{\min} < 0.5C_1^{-1}\} \cup \{\mathcal{D}_n : \hat{\lambda}_{\max} > 1.5C_1\} \\ &\cup \{\mathcal{D}_n : \|\mathbb{E}_n[\phi^{J_n}(W_i)R_i]\|_{\ell^2}^2 > M_n\delta_n^2\}.\end{aligned}$$

Construct the "tests" ω_n by $\omega_n = 1(\mathcal{E}_{1n})$. Then we have

$$\begin{aligned}&\tilde{\Pi}_n(b^{J_n} : \|b^{J_n} - b_0^{J_n}\|_{\mathcal{D}_n} > M_n\delta_n \mid \mathcal{D}_n) \\ &\quad = \tilde{\Pi}_n(b^{J_n} : \|b^{J_n} - b_0^{J_n}\|_{\mathcal{D}_n} > M_n\delta_n \mid \mathcal{D}_n)\{\omega_n + (1 - \omega_n)\} \\ &\quad \le \omega_n + \tilde{\Pi}_n(b^{J_n} : \|b^{J_n} - b_0^{J_n}\|_{\mathcal{D}_n} > M_n\delta_n \mid \mathcal{D}_n)(1 - \omega_n).\end{aligned} \tag{19}$$

By Lemma 1(ii)–(iv), we have $\mathbb{P}(\omega_n = 1) = \mathbb{P}(\mathcal{E}_{1n}) \to 0$.

For the second term in (19), taking $c > 0$ sufficiently small in (17), we have

$$\begin{aligned}&(1 - \omega_n)\int_{\|b^{J_n} - b_0^{J_n}\|_{\mathcal{D}_n} > M_n\delta_n} \exp\{-(n/2)\ell_{b^{J_n}}(\mathcal{D}_n)\}\tilde{\Pi}_n(db^{J_n}) \\ &\quad \le \exp\{-c_1 M_n^2 n\delta_n^2 + O(M_n n\delta_n^2)\} \le e^{-c_2 M_n^2 n\delta_n^2}.\end{aligned}$$

On the other hand, taking, say $c = 1$ in (18), we have

$$\begin{aligned}&(1 - \omega_n)\int \exp\{-(n/2)\ell_{b^{J_n}}(\mathcal{D}_n)\}\tilde{\Pi}_n(db^{J_n}) \\ &\quad \ge (1 - \omega_n)\int_{\|b^{J_n} - b_0^{J_n}\|_{\mathcal{D}_n} \le \sqrt{M_n}\epsilon_n} \exp\{-(n/2)\ell_{b^{J_n}}(\mathcal{D}_n)\}\tilde{\Pi}_n(db^{J_n}) \\ &\quad \ge (1 - \omega_n)e^{-c_3 M_n n\delta_n^2}\int_{\|b^{J_n} - b_0^{J_n}\|_{\mathcal{D}_n} \le \sqrt{M_n}\epsilon_n} \tilde{\Pi}_n(db^{J_n}).\end{aligned}$$

Denote by $\hat{s}_{\max}$ the maximum singular value of the matrix $\hat{\Phi}_{WX}$, so that

$$\|b^{J_n} - b_0^{J_n}\|_{\mathcal{D}_n} \le \hat{s}_{\max}\|b^{J_n} - b_0^{J_n}\|_{\ell^2}.$$

Define the event $\mathcal{E}_{2n} = \{\mathcal{D}_n : \hat{s}_{\max} \le 1.5C_1\}$. By Lemma 1(ii) and (iii), we have $\mathbb{P}(\mathcal{E}_{2n}) \to 1$. Since $M_n \to \infty$, for all n sufficiently large, we have

$$\begin{aligned}
&1(\mathcal{E}_{2n})(1-\omega_n)\int \exp\{-(n/2)\ell_{b^{J_n}}(\mathcal{D}_n)\}\tilde{\Pi}_n(db^{J_n}) \\
&\quad \ge 1(\mathcal{E}_{2n})(1-\omega_n)e^{-c_3 M_n n\delta_n^2}\tilde{\Pi}_n(b^{J_n} : \|b^{J_n} - b_0^{J_n}\|_{\ell 2} \le \epsilon_n) \\
&\quad \ge 1(\mathcal{E}_{2n})(1-\omega_n)e^{-c_3 M_n \delta_n^2 - Cn\epsilon_n^2} \\
&\quad \ge 1(\mathcal{E}_{2n})(1-\omega_n)e^{-c_4 M_n n\delta_n^2},
\end{aligned}$$

where the second inequality is due to the small ball condition (P1). Summarizing, we have

$$\tilde{\Pi}_n(b^{J_n} : \|b^{J_n} - b_0^{J_n}\|_{\mathcal{D}_n} > M_n\delta_n \mid \mathcal{D}_n)(1-\omega_n) \le 1(\mathcal{E}_{2n}^c) + e^{-c_2 M_n^2 n\delta_n^2 + c_4 M_n n\delta_n^2}.$$

Therefore, we obtain (15) for a sufficiently small $c_0 > 0$. □

We are now in position to prove Theorem 1. We will say that a sequence of random variables A_n is *eventually* bounded by another sequence of random variables B_n if $\mathbb{P}(A_n \le B_n) \to 1$ as $n \to \infty$.

PROOF OF THEOREM 1. We first note that by Lemma 1(ii), (iii) and (v), the matrices $\hat{\Phi}_{WX}$ and $\hat{\Phi}_{WW}$ are nonsingular with probability approaching one. Conditional on $\mathcal{D}_n$, define the rescaled "parameter" $\theta^{J_n} = (\theta_1, \ldots, \theta_{2^{J_n}})^T = \sqrt{n}\hat{\Phi}_{WX}(b^{J_n} - b_0^{J_n})$. By (16), the corresponding "quasi-posterior" density for θ^{J_n} is given by

$$\pi_n^*(\theta^{J_n} \mid \mathcal{D}_n)\, d\theta^{J_n} \propto \tilde{\pi}_n(b_0^{J_n} + \hat{\Phi}_{WX}^{-1}\theta^{J_n}/\sqrt{n})\, dN(\Delta_n, \hat{\Phi}_{WW})(\theta^{J_n})\, d\theta^{J_n},$$

where recall that $\Delta_n = \sqrt{n}\mathbb{E}_n[\phi^{J_n}(W_i)R_i]$ (this operation is valid as soon as $\hat{\Phi}_{WX}$ and $\hat{\Phi}_{WW}$ are nonsingular, of which the probability is approaching one).

The proof of Theorem 1 consists of 3 steps. After step 1, we will turn to the proof of (9). The remaining two steps are devoted to the proof of (10).

Step 1. We first show that

$$\int \left|\pi_n^*(\theta^{J_n} \mid \mathcal{D}_n) - dN(\Delta_n, \hat{\Phi}_{WW})(\theta^{J_n})\right| d\theta^{J_n} \xrightarrow{P} 0. \tag{20}$$

In this step, we do *not* assume $J_n 2^{3J_n}/n = o(\tau_{J_n}^2)$. As before, let $\delta_n = \epsilon_n + \tau_{J_n}2^{-J_n s}$. By Proposition 4, for every sequence $M_n \to \infty$,

$$\int_{\|\theta^{J_n}\|_{\ell^2} \le M_n\sqrt{n}\delta_n} \pi_n^*(\theta^{J_n} \mid \mathcal{D}_n)\, d\theta^{J_n} = 1 + o_P(1)$$

by which we have

$$
\begin{aligned}
&\text{Left-hand side of (20)}\\
&\le \int_{\|\theta^{J_n}\|_{\ell^2}\le M_n\sqrt{n}\delta_n} \big|\pi_n^*(\theta^{J_n}\mid \mathcal{D}_n) - dN(\Delta_n,\hat{\Phi}_{WW})(\theta^{J_n})\big|\,d\theta^{J_n} \qquad (21)\\
&\quad + \int_{\|\theta^{J_n}\|_{\ell^2}> M_n\sqrt{n}\delta_n} dN(\Delta_n,\hat{\Phi}_{WW})(\theta^{J_n})\,d\theta^{J_n} + o_P(1).
\end{aligned}
$$

By Lemma 1(iv), $\|\Delta_n\|_{\ell^2} = O_P(\sqrt{n}\delta_n)$, and by Lemma 1(ii) and (iii), $(1 - o_P(1))C_1^{-1} \le s_{\min}(\hat{\Phi}_{WW}) \le s_{\max}(\hat{\Phi}_{WW}) \le (1+o_P(1))C_1$, so that the second integral is eventually bounded by

$$
\int_{\|\theta^{J_n}\|_{\ell^2}>\sqrt{M_n n}\delta_n} dN(0, I_{2^{J_n}})(\theta^{J_n})\,d\theta^{J_n}, \qquad (22)
$$

where note that M_n is replaced by $\sqrt{M_n}$ to "absorb" the constant. By Borell's inequality for Gaussian measures (see, e.g., [55], Lemma A.2.2), for every $x > 0$,

$$
\mathbb{P}\big(\|N(0, I_{2^{J_n}})\|_{\ell^2} > \sqrt{2^{J_n}} + x\big) \le e^{-x^2/2}. \qquad (23)
$$

Here since $n\delta_n^2 \ge n\epsilon_n^2 \gtrsim 2^{J_n}$, $\sqrt{M_n n}\delta_n/\sqrt{2^{J_n}} \to \infty$, so that the integral in (22) is $o(1)$.

It remains to show that the first integral in (21) is $o_P(1)$. This step uses a standard cancellation argument. Let $\mathcal{C}_n := \{\theta^{J_n} \in \mathbb{R}^{2^{J_n}} : \|\theta^{J_n}\|_{\ell^2} \le M_n\sqrt{n}\delta_n\}$. First, provided that $\|\hat{\Phi}_{WX}^{-1}\|_{\mathrm{op}} \le 1.5\tau_{J_n}^{-1}$, for all $\theta^{J_n} \in \mathcal{C}_n$,

$$
\|\hat{\Phi}_{WX}^{-1}\theta^{J_n}/\sqrt{n}\|_{\ell^2} \le 1.5M_n\tau_{J_n}^{-1}\delta_n \le 1.5M_n(2^{-J_n s} + \tau_{J_n}^{-1}\epsilon_n) \sim M_n\gamma_n.
$$

So taking $M_n \to \infty$ such that $M_n = o(L_n)$, $\|\hat{\Phi}_{WX}^{-1}\theta^{J_n}/\sqrt{n}\|_{\ell^2} \le L_n\gamma_n$ and hence $\tilde{\pi}_n(b_0^{J_n} + \hat{\Phi}_{WX}^{-1}\theta^{J_n}/\sqrt{n}) > 0$ for all n sufficiently large. Here, by Lemma 1(v), we have $\mathbb{P}(\|\hat{\Phi}_{WX}^{-1}\|_{\mathrm{op}} \le 1.5\tau_{J_n}^{-1}) \to 1$.

Suppose that $\|\hat{\Phi}_{WX}^{-1}\|_{\mathrm{op}} \le 1.5\tau_{J_n}^{-1}$. Let

$$
\pi_{n,\mathcal{C}_n}^*(\theta^{J_n}\mid \mathcal{D}_n) \quad \text{and} \quad dN^{\mathcal{C}_n}(\Delta_n,\hat{\Phi}_{WW})(\theta^{J_n})
$$

denote the probability densities obtained by first restricting $\pi_n^*(\theta^{J_n}\mid\mathcal{D}_n)$ and $dN(\Delta_n,\hat{\Phi}_{WW})(\theta^{J_n})$ to the ball $\mathcal{C}_n$ and then renormalizing, respectively. By the first part of the present proof, replacing $\pi_n^*(\theta^{J_n}\mid\mathcal{D}_n)$ and $dN(\Delta_n,\hat{\Phi}_{WW})(\theta^{J_n})$ by $\pi_{n,\mathcal{C}_n}^*(\theta^{J_n}\mid\mathcal{D}_n)$ and $dN^{\mathcal{C}_n}(\Delta_n,\hat{\Phi}_{WW})(\theta^{J_n})$, respectively, in the first integral in (21) has impact at most $o_P(1)$. Abbreviating $\pi_{n,\mathcal{C}_n}^*(\theta^{J_n}\mid\mathcal{D}_n)$ by $\pi_{n,\mathcal{C}_n}^*$, $dN^{\mathcal{C}_n}(\Delta_n,\hat{\Phi}_{WW})(\theta^{J_n})$ by $dN^{\mathcal{C}_n}$, $dN(\Delta_n,\hat{\Phi}_{WW})(\theta^{J_n})$ by dN, and $\tilde{\pi}_n(b_0^{J_n} +$

$\hat{\Phi}_{WX}^{-1}\theta^{J_n}/\sqrt{n})$ by $\tilde{\pi}_n$, we have

$$\int |\pi^*_{n,\mathcal{C}_n} - dN^{\mathcal{C}_n}| = \int \left|1 - \frac{dN^{\mathcal{C}_n}}{\pi^*_{n,\mathcal{C}_n}}\right| \pi^*_{n,\mathcal{C}_n} = \int \left|1 - \frac{dN/\int_{\mathcal{C}_n} dN}{\tilde{\pi}_n dN/\int_{\mathcal{C}_n} \tilde{\pi}_n dN}\right| \pi^*_{n,\mathcal{C}_n}$$
$$= \int \left|1 - \frac{\int_{\mathcal{C}_n} \tilde{\pi}_n dN}{\tilde{\pi}_n \int_{\mathcal{C}_n} dN}\right| \pi^*_{n,\mathcal{C}_n} = \int \left|1 - \frac{\int_{\mathcal{C}_n} \tilde{\pi}_n dN^{\mathcal{C}_n}}{\tilde{\pi}_n}\right| \pi^*_{n,\mathcal{C}_n}.$$

By the convexity of the map $x \mapsto |1-x|$ and Jensen's inequality, the last expression is bounded by

$$\sup_{\theta^{J_n} \in \mathcal{C}_n, \tilde{\theta}^{J_n} \in \mathcal{C}_n} \left|1 - \frac{\tilde{\pi}_n(b_0^{J_n} + \hat{\Phi}_{WX}^{-1}\theta^{J_n}/\sqrt{n})}{\tilde{\pi}_n(b_0^{J_n} + \hat{\Phi}_{WX}^{-1}\tilde{\theta}^{J_n}/\sqrt{n})}\right|,$$

which is eventually bounded by

$$\sup_{\|b^{J_n}\|_{\ell^2} \le L_n\gamma_n, \|\tilde{b}^{J_n}\|_{\ell^2} \le L_n\gamma_n} \left|1 - \frac{\tilde{\pi}_n(b_0^{J_n} + b^{J_n})}{\tilde{\pi}_n(b_0^{J_n} + \tilde{b}^{J_n})}\right|.$$

The last expression goes to zeros as $n \to \infty$ by condition (P2).

We now turn to the proof of (9). Take any $M_n \to \infty$ (this M_n may be different from the previous M_n). By step 1, we have

$$\sup_{z>0}\Big| \tilde{\Pi}_n\{b^{J_n} : \|\hat{\Phi}_{WX}(b^{J_n} - b_0^{J_n})\|_{\ell^2} > z \mid \mathcal{D}_n\}$$
$$- \int_{\|\theta^{J_n}\|_{\ell^2} > z} dN(n^{-1/2}\Delta_n, n^{-1}\hat{\Phi}_{WW})(\theta^{J_n})\, d\theta^{J_n}\Big| \xrightarrow{P} 0.$$

By Lemma 1(v), we have

$$\begin{aligned}\|\hat{\Phi}_{WX}(b^{J_n} - b_0^{J_n})\|_{\ell^2} &\ge s_{\min}(\hat{\Phi}_{WX})\|b^{J_n} - b_0^{J_n}\|_{\ell^2} \\ &\ge (1 - o_P(1))\tau_{J_n}\|b^{J_n} - b_0^{J_n}\|_{\ell^2}\end{aligned}$$

by which we have, uniformly in $z > 0$,

$$\begin{aligned}&\tilde{\Pi}_n\{b^{J_n} : \|b^{J_n} - b_0^{J_n}\|_{\ell^2} > 2\tau_{J_n}^{-1}z \mid \mathcal{D}_n\} \\ &\quad \le \tilde{\Pi}_n\{b^{J_n} : \|\hat{\Phi}_{WX}(b^{J_n} - b_0^{J_n})\|_{\ell^2} > z \mid \mathcal{D}_n\} + o_P(1) \\ &\quad \le \int_{\|\theta^{J_n}\|_{\ell^2} > z} dN(n^{-1/2}\Delta_n, n^{-1}\hat{\Phi}_{WW})(\theta^{J_n})\, d\theta^{J_n} + o_P(1).\end{aligned}$$

By Markov's inequality, the integral in the last expression is bounded by

$$\frac{1}{nz^2}\{\|\Delta_n\|_{\ell^2}^2 + \mathrm{tr}(\hat{\Phi}_{WW})\}.$$

By Lemma 1(ii)–(iv), we have $\|\Delta_n\|_{\ell^2}^2 + \mathrm{tr}(\hat{\Phi}_{WW}) = O_P(2^{J_n} + n\tau_{J_n}^2 2^{-2J_n s})$. Therefore, we conclude that, taking $z = M_n(\tau_{J_n} 2^{-J_n s} + \sqrt{2^{J_n}/n})$, $\tilde{\Pi}_n\{b^{J_n} : \|b^{J_n} - b_0^{J_n}\|_{\ell^2} > 2M_n(2^{-J_n s} + \tau_{J_n}^{-1}\sqrt{2^{J_n}/n}) \mid \mathcal{D}_n\} \xrightarrow{P} 0$, which leads to the contraction rate result (9).

In what follows, we assume $J_n 2^{3J_n}/n = o(\tau_{J_n}^2)$, and prove the asymptotic normality result (10).

Step 2 (replacement of $\hat{\Phi}_{WW}$ by Φ_{WW}). This step shows that

$$\int \left| dN(\Delta_n, \hat{\Phi}_{WW})(\theta^{J_n}) - dN(\Delta_n, \Phi_{WW})(\theta^{J_n}) \right| d\theta^{J_n} \xrightarrow{P} 0,$$

which is equivalent to

$$\int \left| dN(0, \hat{\Phi}_{WW})(\theta^{J_n}) - dN(0, \Phi_{WW})(\theta^{J_n}) \right| d\theta^{J_n} \xrightarrow{P} 0.$$

By Lemmas 1(ii), (iii) and 2, this follows if $\sqrt{J_n 2^{J_n}/n} = o(2^{-J_n})$, that is, $J_n 2^{3J_n} = o(n)$, which is satisfied since $J_n 2^{3J_n}/n = o(\tau_{J_n}^2) = o(1)$.

Step 3 (replacement of $\hat{\Phi}_{WX}$ by Φ_{WX}). We have shown that

$$\int \left| \pi_n^*(\theta^{J_n} \mid \mathcal{D}_n) - dN(\Delta_n, \Phi_{WW})(\theta^{J_n}) \right| d\theta^{J_n} \xrightarrow{P} 0.$$

By Scheffé's lemma, this means that

$$\left\| \tilde{\Pi}_n\{b^{J_n} : \sqrt{n}\hat{\Phi}_{WX}(b^{J_n} - b_0^{J_n}) \in \cdot \mid \mathcal{D}_n\} - N(\Delta_n, \Phi_{WW})(\cdot) \right\|_{\mathrm{TV}} \xrightarrow{P} 0$$

or equivalently,

$$\left\| \tilde{\Pi}_n\{b^{J_n} : \sqrt{n}(b^{J_n} - b_0^{J_n}) \in \cdot \mid \mathcal{D}_n\} - N(\hat{\Phi}_{WX}^{-1}\Delta_n, \hat{\Phi}_{WX}^{-1}\Phi_{WW}\hat{\Phi}_{XW}^{-1})(\cdot) \right\|_{\mathrm{TV}} \xrightarrow{P} 0.$$

The last expression is asymptotically valid since $\hat{\Phi}_{WX}$ is nonsingular with probability approaching one. Recall the maximum quasi-likelihood estimator $\hat{b}^{J_n}$. With probability approaching one, we have

$$\hat{b}^{J_n} = \hat{\Phi}_{WX}^{-1}\mathbb{E}_n[\phi^{J_n}(W_i)Y_i] = b_0^{J_n} + \hat{\Phi}_{WX}^{-1}\mathbb{E}_n[\phi^{J_n}(W_i)R_i],$$

so that $\sqrt{n}(\hat{b}^{J_n} - b_0^{J_n}) = \hat{\Phi}_{WX}^{-1}\Delta_n$. Hence to conclude the theorem, it suffices to show that

$$\begin{aligned} &\left\| N(\hat{\Phi}_{WX}^{-1}\Delta_n, \hat{\Phi}_{WX}^{-1}\Phi_{WW}\hat{\Phi}_{XW}^{-1}) \right. \\ &\qquad \left. - N(\hat{\Phi}_{WX}^{-1}\Delta_n, \Phi_{WX}^{-1}\Phi_{WW}\Phi_{XW}^{-1}) \right\|_{\mathrm{TV}} \xrightarrow{P} 0. \end{aligned} \tag{24}$$

Assertion (24) reduces to

$$\left\| N(0, \Phi_{WX}\hat{\Phi}_{WX}^{-1}\Phi_{WW}\hat{\Phi}_{XW}^{-1}\Phi_{XW}) - N(0, \Phi_{WW}) \right\|_{\mathrm{TV}} \xrightarrow{P} 0.$$

By Lemmas 1(ii), (iii) and 3,

$$\|\Phi_{WX}\hat{\Phi}_{WX}^{-1}\Phi_{WW}\hat{\Phi}_{XW}^{-1}\Phi_{XW}-\Phi_{WW}\|_{\mathrm{op}}=O_P\big(\tau_{J_n}^{-1}\sqrt{J_n 2^{J_n}/n}\big)=o_P\big(2^{-J_n}\big)$$

[the last equality follows since $J_n 2^{3J_n}/n=o(\tau_{J_n}^2)$]. Since $C_1^{-1}\le s_{\min}(\Phi_{WW})\le s_{\max}(\Phi_{WW})\le C_1$, the desired conclusion follows from Lemma 2.

Steps 1–3 lead to the asymptotic normality result (10). □

5.2. *Proof of Theorem* 2. We first prove the following lemma.

LEMMA 4. *Suppose that the conditions of Theorem* 2 *are satisfied. Then there exists a constant* $D>0$ *such that*

$$\mathbb{P}\{\|\mathbb{E}_n[\phi^{J_n}(W_i)U_i]\|_{\ell^2}>D\sqrt{2^{J_n}/n}\}\to 0.$$

REMARK 9. It is standard to show that $\|\mathbb{E}_n[\phi^{J_n}(W_i)U_i]\|_{\ell^2}=O_P(\sqrt{2^{J_n}/n})$, which, however, does not leads to the conclusion of Lemma 4 since the former only implies that for every sequence $M_n\to\infty$, $\mathbb{P}\{\|\mathbb{E}_n[\phi^{J_n}(W_i)U_i]\|_{\ell^2}>M_n\sqrt{2^{J_n}/n}\}\to 0$. Hence, an additional step is needed. The current proof uses a truncation argument and Talagrand's concentration inequality.

PROOF OF LEMMA 4. For a given $\lambda>0$, define $U_i^-=U_i 1(|U_i|\le\lambda)$ and $U_i^+=U_i 1(|U_i|>\lambda)$. Since $0=\mathbb{E}[U\mid W]=\mathbb{E}[U^-\mid W]+\mathbb{E}[U^+\mid W]$, we have $\mathbb{E}_n[\phi^{J_n}(W_i)U_i]=n^{-1}\sum_{i=1}^n\{\phi^{J_n}(W_i)U_i^- - \mathbb{E}[\phi^{J_n}(W)U^-]\}+n^{-1}\times\sum_{i=1}^n\{\phi^{J_n}(W_i)U_i^+ - \mathbb{E}[\phi^{J_n}(W)U^+]\}$, by which we have

$$\begin{aligned}\|\mathbb{E}_n[\phi^{J_n}(W_i)U_i]\|_{\ell^2}\le{}&\left\|n^{-1}\sum_{i=1}^n\{\phi^{J_n}(W_i)U_i^- - \mathbb{E}[\phi^{J_n}(W)U^-]\}\right\|_{\ell^2}\\&+\left\|n^{-1}\sum_{i=1}^n\{\phi^{J_n}(W_i)U_i^+ - \mathbb{E}[\phi^{J_n}(W)U^+]\}\right\|_{\ell^2}\\&=: I+II.\end{aligned}$$

First, by Markov's inequality, we have for every $z>0$,

$$\begin{aligned}\mathbb{P}(II>z)&\le\frac{\mathbb{E}[II^2]}{z^2}\le\frac{\sum_{l=1}^{2^{J_n}}\mathbb{E}[(\phi_l(W)U^+)^2]}{nz^2}\\&\le\frac{\sup_{w\in[0,1]}\mathbb{E}[U^2 1(|U|>\lambda)\mid W=w]\times\sum_{l=1}^{2^{J_n}}\mathbb{E}[\phi_l(W)^2]}{nz^2}\\&\le\frac{C_1 2^{J_n}}{nz^2}\times\sup_{w\in[0,1]}\mathbb{E}\big[U^2 1\big(|U|>\lambda\big)\mid W=w\big],\end{aligned}$$

where we have used that $\sum_{l=1}^{2^{J_n}} \mathbb{E}[\phi_l(W)^2] = \mathrm{tr}(\Phi_{WW}) \le 2^{J_n} s_{\max}(\Phi_{WW}) \le C_1 2^{J_n}$ by Lemma 1(ii). Thus, we have

$$\mathbb{P}\{II > \sqrt{C_1 2^{J_n}/n}\} \le \sup_{w\in[0,1]} \mathbb{E}[U^2 1(|U| > \lambda) \mid W = w].$$

By assumption, the right-hand side goes to zero as $\lambda \to \infty$.

Second, let $Z_i = \phi^{J_n}(W_i)U_i^- - \mathbb{E}[\phi^{J_n}(W)U^-]$ (denote by Z the generic version of Z_i). Let $\mathbb{S}^{2^{J_n}-1} := \{\alpha^{J_n} \in \mathbb{R}^{2^{J_n}} : \|\alpha^{J_n}\|_{\ell^2} = 1\}$. Then

$$I = \|\mathbb{E}_n[Z_i]\|_{\ell^2} = \sup_{\alpha^{J_n}\in\mathbb{S}^{2^{J_n}-1}} \mathbb{E}_n[(\alpha^{J_n})^T Z_i].$$

We make use of Talagrand's concentration inequality to bound the tail probability of I. For any $\alpha^{J_n} \in \mathbb{S}^{2^{J_n}-1}$, by Lemma 1, we have

$$\mathbb{E}[\{(\alpha^{J_n})^T Z\}^2] \le \sup_{w\in[0,1]} \mathbb{E}[U^2 \mid W = w] \times s_{\max}(\Phi_{WW}) \le C_1^2,$$

$$|(\alpha^{J_n})^T Z| \le \lambda \sup_{w\in[0,1]} \|\phi^{J_n}(w)\|_{\ell^2} \le D_1\lambda\sqrt{2^{J_n}}$$

and

$$\begin{aligned}(\mathbb{E}[I])^2 \le \mathbb{E}[I^2] &\le n^{-1} \sup_{w\in[0,1]} \mathbb{E}[U^2 \mid W = w] \times \sum_{l=1}^{2^{J_n}} \mathbb{E}[\phi_l(W)^2] \\ &\le C_1^2 2^{J_n}/n,\end{aligned}$$

where $D_1 > 0$ is a constant. Thus, by Talagrand's inequality (see Theorem 2 in Appendix E), we have for every $z > 0$

$$\mathbb{P}\{I \ge D_2(\sqrt{2^{J_n}/n} + \sqrt{z/n} + z\lambda\sqrt{2^{J_n}}/n)\} \le e^{-z},$$

where $D_2 > 0$ is a constant independent of λ and z.

The final conclusion follows from taking $\lambda = \lambda_n \to \infty$ and $z = z_n \to \infty$ sufficiently slowly. $\square$

PROOF OF THEOREM 2. Let D_1 and D_2 be some positive constants of which the values are understood in the context. For either $g_0 \in B^s_{\infty,\infty}$ or $B^s_{2,2}$, $\|g_0 - P_{J_n} g_0\| = O(2^{-J_n s}) = o(1)$, by which we have

$$\begin{aligned}&\sum_{l=1}^{2^{J_n}} \mathrm{Var}\{\mathbb{E}_n[\phi_l(W_i)(g_0 - P_{J_n} g_0)(X_i)]\} \\ &\quad \le n^{-1} \sum_{l=1}^{2^{J_n}} \mathbb{E}[\phi_l(W)^2\{(g_0 - P_{J_n} g_0)(X)\}^2]\end{aligned}$$

$$
\begin{aligned}
&= n^{-1} \sum_{l=1}^{2^{J_n}} \iint \phi_l(w)^2 \{(g_0 - P_{J_n} g_0)(x)\}^2 f_{X,W}(x,w)\,dx\,dw \\
&\le n^{-1} C_1 \|g_0 - P_{J_n} g_0\|^2 \times \sum_{l=1}^{2^{J_n}} \int \phi_l(w)^2\,dw = o(2^{J_n}/n).
\end{aligned}
$$

Hence

$$
\begin{aligned}
\mathbb{E}_n[\phi^{J_n}(W_i)R_i] = \mathbb{E}_n[\phi^{J_n}(W_i)U_i] \\
+ \mathbb{E}[\phi^{J_n}(W)(g_0 - P_n g_0)(X)] + \mathrm{Rem}
\end{aligned}
$$

with $\|\mathrm{Rem}\|_{\ell^2} = o_P(\sqrt{2^{J_n}/n})$. The second term on the right-hand side is $O(\tau_{J_n} 2^{-J_n s})$ in the Euclidean norm. Together with Lemma 4, we have

$$
\mathbb{P}\{\|\mathbb{E}_n[\phi^{J_n}(W_i)R_i]\|_{\ell^2}^2 > D_1(\tau_{J_n}^2 2^{-2J_n s} + 2^{J_n}/n)\} \to 0.
$$

Moreover, by Lemma 1, we have

$$
\mathrm{tr}(\hat{\Phi}_{WW}) \le 2^{J_n} s_{\max}(\hat{\Phi}_{WW}) \le C_1(1 + o_P(1))2^{J_n}.
$$

Taking these together, we have

$$
\mathbb{P}\{\|\mathbb{E}_n[\phi^{J_n}(W_i)R_i]\|_{\ell^2}^2 + n^{-1}\,\mathrm{tr}(\hat{\Phi}_{WW}) \le D_2(\tau_{J_n}^2 2^{-2J_n s} + 2^{J_n}/n)\} \to 1.
$$

By the proof of Theorem 1, this leads to the desired conclusion. □

6. Discussion. We have studied the asymptotic properties of quasi-posterior distributions against sieve priors in the NPIV model and given some specific priors for which the quasi-posterior distribution (the quasi-Bayes estimator) attains the minimax optimal rate of contraction (convergence, resp.). These results greatly sharpen the previous work [44]. We end this paper with two additional discussions.

6.1. *Multivariate case.* In this paper, we have focused on the case where X and W are scalar, mainly to avoid the notational complication. It is not difficult to see that the results naturally extend to the case where X and W are vectors with the same dimension, by considering tensor product sieves (the contraction/convergence rates will then deteriorate as the dimension grows). We can also consider the following more general situation as in Section 3 of [28]: suppose that Y is a scalar random variable, X and W are random vectors with the same dimension, and Z is another random vector (whose dimension may be different from X), and suppose that we are interested in estimating the function g_0 identified by the conditional moment restriction: $\mathbb{E}[Y \mid Z, W] = \mathbb{E}[g_0(X,Z) \mid Z, W]$ or $Y = g_0(X,Z) + U$ with $\mathbb{E}[U \mid Z, W] = 0$ (i.e., X and Z are endogenous and exogenous explanatory variables, resp.). In principle, the analysis can be reduced to the case where there are no exogenous variables by conditioning on $Z = z$ (so the sieve measure of ill-posedness can be defined by the one conditional on $Z = z$). More precisely, when Z is discretely distributed with finitely many mass points, then $g_0(x, z)$, where z is a mass point, can be estimated by using only observations i for which $Z_i = z$. When Z is continuously distributed, then $g_0(x, z)$ can be estimated by using observations i for which Z_i is "close" to z; one way is to use kernel weights as in Section 4.2 of [31]. However, the detailed analysis of this case is not presented here for brevity.

6.2. *Direction of future research.* Finally, we make some remarks on the direction of future research. First, as also noted by [44], (adaptive) selection of the resolution level J_n in a (quasi-)Bayesian or "empirical" Bayesian approach is an important topic to be investigated. Second, a (quasi-)Bayesian analysis is typically useful in the analysis of complex models in which frequentist estimation is difficult to implement due to nondifferentiability/nonconvex nature of loss functions. This usefulness comes from the fact that a (quasi-)Bayesian approach is typically able to avoid numerical optimization. See [12] and [53] for the finite-dimensional case. In infinite-dimensional models, such a computational challenge in frequentist estimation occurs in the analysis of nonparametric instrumental quantile regression models [10, 21, 33]. In that model, a typical loss function contains the indicator function and hence highly nonconvex. In such a case, the computation of an optimal solution is by itself difficult, and a solution obtained, if possible, is typically not guaranteed to be globally optimal since there may be many local optima. It is hence of interest to extend the results of the paper to nonparametric instrumental quantile regression models. The extension to the quantile regression case, which is currently under investigation, is highly nontrivial since the problem of estimating the structural function becomes a *nonlinear* ill-posed inverse problem and a delicate care of the stochastic expansion of the criterion function is needed.

Acknowledgments. A major part of the work was done while the author was visiting the Department of Economics, MIT. He would like to thank Professor Victor Chernozhukov for his suggestions and encouragements, as well as Professor Yukitoshi Matsushita for his constructive comments. Also he would like to thank the Editor, Professor Runze Li, the Associate Editor, and anonymous referees for their insightful comments that helped improve on the quality of the paper.

SUPPLEMENTARY MATERIAL

Supplement to "Quasi-Bayesian analysis of nonparametric instrumental variables models" (DOI: 10.1214/13-AOS1150SUPP; .pdf). This supplemental file contains the additional technical proofs omitted in the main text, and some technical tools used in the proofs.

REFERENCES

[1] AGAPIOU, S., LARSSON, S. and STUART, A. M. (2013). Posterior contraction rates for the Bayesian approach to linear ill-posed inverse problems. *Stochastic Process. Appl.* **123** 3828–3860.

[2] AI, C. and CHEN, X. (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica* **71** 1795–1843. MR2015420

[3] BELLONI, A. and CHERNOZHUKOV, V. (2009). Posterior inference in curved exponential families under increasing dimensions. Preprint. Available at arXiv:0904.3132.

[4] BELLONI, A. and CHERNOZHUKOV, V. (2009). On the computational complexity of MCMC-based estimators in large samples. *Ann. Statist.* **37** 2011–2055. MR2533478

[5] BLUNDELL, R., CHEN, X. and KRISTENSEN, D. (2007). Semi-nonparametric IV estimation of shape-invariant Engel curves. *Econometrica* **75** 1613–1669. MR2351452

[6] BONTEMPS, D. (2011). Bernstein–von Mises theorems for Gaussian regression with increasing number of regressors. *Ann. Statist.* **39** 2557–2584. MR2906878

[7] BOUCHERON, S. and GASSIAT, E. (2009). A Bernstein–von Mises theorem for discrete probability distributions. *Electron. J. Stat.* **3** 114–148. MR2471588

[8] CARRASCO, M., FLORENS, J. P. and RENAULT, E. (2007). Linear inverse problems in structural econometrics: Estimation based on spectral decomposition and regularization. In *Handbook of Econometrics*, *Vol.* 6 (J. J. Heckman and E. E. Leamer, eds.) 5633–5751. Elsevier, Amsterdam.

[9] CAVALIER, L. (2008). Nonparametric statistical inverse problems. *Inverse Problems* **24** 034004, 19. MR2421941

[10] CHEN, X. and POUZO, D. (2012). Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. *Econometrica* **80** 277–321. MR2920758

[11] CHEN, X. and REISS, M. (2011). On rate optimality for ill-posed inverse problems in econometrics. *Econometric Theory* **27** 497–521. MR2806258

[12] CHERNOZHUKOV, V. and HONG, H. (2003). An MCMC approach to classical estimation. *J. Econometrics* **115** 293–346. MR1984779

[13] CLARKE, B. and GHOSAL, S. (2010). Reference priors for exponential families with increasing dimension. *Electron. J. Stat.* **4** 737–780. MR2678969

[14] COHEN, A., DAUBECHIES, I. and VIAL, P. (1993). Wavelets on the interval and fast wavelet transforms. *Appl. Comput. Harmon. Anal.* **1** 54–81. MR1256527

[15] COX, D. D. (1993). An analysis of Bayesian inference for nonparametric regression. *Ann. Statist.* **21** 903–923. MR1232525

[16] DAROLLES, S., FAN, Y., FLORENS, J. P. and RENAULT, E. (2011). Nonparametric instrumental regression. *Econometrica* **79** 1541–1565. MR2883763

[17] D'HAULTFOEUILLE, X. (2011). On the completeness condition in nonparametric instrumental problems. *Econometric Theory* **27** 460–471. MR2806256

[18] FLORENS, J. P. and SIMONI, A. (2010). Regularized priors for linear inverse problems. IDEI Working Paper 767-2013.

[19] FLORENS, J.-P. and SIMONI, A. (2012). Regularized posteriors in linear ill-posed inverse problems. *Scand. J. Stat.* **39** 214–235. MR2927022

[20] FLORENS, J.-P. and SIMONI, A. (2012). Nonparametric estimation of an instrumental regression: A quasi-Bayesian approach based on regularized posterior. *J. Econometrics* **170** 458–475. MR2970326

[21] GAGLIARDINI, P. and SCAILLET, O. (2012). Nonparametric instrumental variable estimation of structural quantile effects. *Econometrica* **80** 1533–1562. MR2977430

[22] GHOSAL, S. (1999). Asymptotic normality of posterior distributions in high-dimensional linear models. *Bernoulli* **5** 315–331. MR1681701

[23] GHOSAL, S. (2000). Asymptotic normality of posterior distributions for exponential families when the number of parameters tends to infinity. *J. Multivariate Anal.* **74** 49–68. MR1790613
[24] GHOSAL, S., GHOSH, J. K. and VAN DER VAART, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28** 500–531. MR1790007
[25] 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
[26] GHOSH, J. K. and RAMAMOORTHI, R. V. (2003). *Bayesian Nonparametrics*. Springer, New York. MR1992245
[27] GINÉ, E. and NICKL, R. (2011). Rates on contraction for posterior distributions in L^r-metrics, $1 \le r \le \infty$. *Ann. Statist.* **39** 2883–2911. MR3012395
[28] HALL, P. and HOROWITZ, J. L. (2005). Nonparametric methods for inference in the presence of instrumental variables. *Ann. Statist.* **33** 2904–2929. MR2253107
[29] HOFINGER, A. and PIKKARAINEN, H. K. (2007). Convergence rate for the Bayesian approach to linear inverse problems. *Inverse Problems* **23** 2469–2484. MR2441014
[30] HOFINGER, A. and PIKKARAINEN, H. K. (2009). Convergence rates for linear inverse problems in the presence of an additive normal noise. *Stoch. Anal. Appl.* **27** 240–257. MR2503292
[31] HOROWITZ, J. L. (2011). Applied nonparametric instrumental variables estimation. *Econometrica* **79** 347–394. MR2809374
[32] HOROWITZ, J. L. (2012). Specification testing in nonparametric instrumental variable estimation. *J. Econometrics* **167** 383–396. MR2892083
[33] HOROWITZ, J. L. and LEE, S. (2007). Nonparametric instrumental variables estimation of a quantile regression model. *Econometrica* **75** 1191–1208. MR2333498
[34] HU, Y. and SHIU, J. L. (2011). Nonparametric identification using instrumental variables: Sufficient conditions for completeness. Preprint.
[35] 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
[36] KATO, K. (2013). Supplement to "Quasi-Bayesian analysis of nonparametric instrumental variables models". DOI:10.1214/13-AOS1150SUPP.
[37] KIM, J.-Y. (2002). Limited information likelihood and Bayesian analysis. *J. Econometrics* **107** 175–193. Information and entropy econometrics. MR1889958
[38] KLEIJN, B. J. K. and VAN DER VAART, A. W. (2006). Misspecification in infinite-dimensional Bayesian statistics. *Ann. Statist.* **34** 837–877. MR2283395
[39] KNAPIK, B. T., SZABÓ, B. T., VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2012). Bayes procedures for adaptive inference in inverse problems for the white noise model. Available at arXiv:1209.3628.
[40] KNAPIK, B. T., VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2011). Bayesian inverse problems with Gaussian priors. *Ann. Statist.* **39** 2626–2657. MR2906881
[41] KNAPIK, B. T., VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2013). Bayesian recovery of the initial condition for the heat equation. *Comm. Statist. Theory Methods* **42** 1294–1313. MR3031282
[42] KOVCHEGOV, Y. and YILDIZ, N. (2012). Identification via completeness for discrete covariates and orthogonal polynomials. Preprint.
[43] KRESS, R. (1999). *Linear Integral Equations*, 2nd ed. *Applied Mathematical Sciences* **82**. Springer, New York. MR1723850
[44] LIAO, Y. and JIANG, W. (2011). Posterior consistency of nonparametric conditional moment restricted models. *Ann. Statist.* **39** 3003–3031. MR3012399
[45] NEWEY, W. K. and POWELL, J. L. (2003). Instrumental variable estimation of nonparametric models. *Econometrica* **71** 1565–1578. MR2000257

[46] SANTOS, A. (2012). Inference in nonparametric instrumental variables with partial identification. *Econometrica* **80** 213–275. MR2920757

[47] SCHENNACH, S. M. (2005). Bayesian exponentially tilted empirical likelihood. *Biometrika* **92** 31–46. MR2158608

[48] SCRICCIOLO, C. (2006). Convergence rates for Bayesian density estimation of infinite-dimensional exponential families. *Ann. Statist.* **34** 2897–2920. MR2329472

[49] SEVERINI, T. A. and TRIPATHI, G. (2006). Some identification issues in nonparametric linear models with endogenous regressors. *Econometric Theory* **22** 258–278. MR2230389

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

[51] STUART, A. M. (2010). Inverse problems: A Bayesian perspective. *Acta Numer.* **19** 451–559. MR2652785

[52] TALAGRAND, M. (1996). New concentration inequalities in product spaces. *Invent. Math.* **126** 505–563. MR1419006

[53] TIAN, L., LIU, J. S. and WEI, L. J. (2007). Implementation of estimating function-based inference procedures with Markov chain Monte Carlo samplers. *J. Amer. Statist. Assoc.* **102** 881–888. MR2411651

[54] VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247

[55] VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer, New York. With applications to statistics. MR1385671

[56] YOSIDA, K. (1980). *Functional Analysis*, 6th ed. *Grundlehren der Mathematischen Wissenschaften* [*Fundamental Principles of Mathematical Sciences*] **123**. Springer, Berlin. MR0617913

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

[58] ZHANG, T. (2006). Information-theoretic upper and lower bounds for statistical estimation. *IEEE Trans. Inform. Theory* **52** 1307–1321. MR2241190

GRADUATE SCHOOL OF ECONOMICS
UNIVERSITY OF TOKYO
7-3-1 HONGO, BUNKYO-KU
TOKYO 113-0033
JAPAN
E-MAIL: kkato@e.u-tokyo.ac.jp