BAYESIAN INFERENCE ON MONOTONE REGRESSION QUANTILE: COVERAGE AND RATE ACCELERATION

BY MOUMITA CHAKRABORTY AND SUBHASISH GHOSAL*

North Carolina State University

For a univariate monotone regression function \( f \), the location \( f^{-1}(y_0) \) where a value \( y_0 \) is attained, is called a regression quantile. We study the coverage of a Bayesian credible interval in a nonparametric monotone regression model for \( f^{-1}(y_0) \), assuming that it is unique and the regression function has a positive derivative there. We put a prior on a piecewise constant function \( f \) with equal intervals and independent normal priors on the step-heights. To comply with the monotonicity constraint for \( f \), we induce a “projection-posterior” by imposing the monotonicity constraint on samples from the posterior distribution of the step-heights. We demonstrate two different interesting phenomena in this context. First, we show that the asymptotic coverage of a credible interval can be higher than the credibility but targeted asymptotic coverage may be obtained by using an appropriate lower credibility level. Next, we show that the posterior contraction rate for the regression quantile can be improved from the optimal \( n^{-1/3} \) level to \( n^{-1/2} \) by sampling in two stages, sparing a fraction of the sampling budget to sample later from a credible interval obtained in the first stage. The first property is analogous to that for the value of the monotone regression function at a point noted by Chakraborty and Ghosal [11] and the second is a Bayesian analog of a similar property for a regression quantile using a frequentist approach by Tang et al. [41].

1. Introduction. Functions appearing in nonparametric modeling such as a regression function or a density function may have shape restrictions like monotonicity or convexity; see Groeneboom and Jongbloed [22] for natural examples. One of the first results was obtained by Grenander [20], who characterized the nonparametric maximum likelihood estimator (MLE) for a decreasing density at any point as the slope of the least concave majorant of the empirical distribution function at that point. Its asymptotic distribution was obtained by Prakasa Rao [34] and identified as the Chernoff distribution (Chernoff [15]), which is given by the probability distribution of

*Research is partially supported by NSF grant number DMS-1916419.


Keywords and phrases: Nonparametric regression, monotonicity, credible set, coverage, Chernoff’s distribution, projection-posterior, two-stage sampling, rate acceleration.
\[ Z = \arg \min \{ W(t) + t^2 : t \in \mathbb{R} \}, \]

where \( W \) is a two-sided standard Brownian motion on \( \mathbb{R} \) starting at 0. The density of \( Z \) and quantiles of its distributions are respectively obtained by Groeneboom [21] and Groeneboom and Wellner [24]. Regression under monotonicity restriction is commonly known as isotonic regression and the MLE is similarly characterized in terms of the slope of the greatest convex minorant (GCM) of a cumulative sum diagram and computed by the Pool-Adjacent-Violators Algorithm (PAVA); see Barlow and Brunk [4] and Ayer et al. [1] for details. Other problems of shape-restricted inference were addressed by Groeneboom and Wellner [23] Huang and Zhang [27], Huang and Wellner [26], and others under various settings. Bayesian approaches to estimating functions under monotonicity were studied by Shivley [38], Salomond [35], and Chakraborty and Ghosal [11, 12, 13]. Tests for the hypothesis of monotonicity were considered by Bowman et al. [7], Hall and Heckman [25], Ghosal et al. [19], and others using the frequentist approach, and by Salomond [36] and Chakraborty and Ghosal [12, 13] using the Bayesian approach.

An exceptionally difficult problem is constructing confidence intervals for the value of a regression or density function with intended coverage and optimal size. This is because under the optimal smoothing, the orders of the squared bias and the variance match, which shifts the limit distribution and hence leads to undercoverage. In parametric problems, this issue does not arise as the bias is much smaller. Confidence sets in shape-restricted inference were constructed by Dümbgen [17], Cai et al. [8], Dümbgen and Johns [18] and Schmidt-Hieber et al. [37]. Banerjee and Wellner [3] and Banerjee [2] proposed a nuisance parameter-free method based on the inversion of the critical region of the likelihood ratio test for testing \( H_0 : f(x_0) = \theta_0 \) against \( H_1 : f(x_0) \neq \theta_0 \). The Bayesian approach to quantifying uncertainty is conceptually simpler as it directly describes variation through a distribution, and an interval for plausible values may be easily obtained by posterior sampling. To justify the construction using a frequentist yardstick, one needs to probabilistically control the size of the interval and show guaranteed asymptotic coverage. However, it was observed by Cox [16] in a Gaussian sequence model that Bayesian credible sets can have arbitrarily low asymptotic coverage for all true functions in the support of the prior. The reason behind this “Cox phenomenon” became clearer through the works of Leahu [31] and Knapik et al. [29]. They also provided sufficient conditions for adequate asymptotic coverage. Other positive coverage results were obtained by Castillo and Nickl [9, 10], Szabó et al. [40], Sniekers and van der Vaart [39], Yoo and Ghosal [43], and Belitser [5] under various settings. Interestingly, the opposite of the Cox phenomenon has been observed recently for
the value of a monotone regression or density function by Chakraborty and Ghosal [11, 13] in that the asymptotic coverage may exceed the credibility level used to construct the interval, while the optimal order of the size is maintained.

In this paper, we consider the nonparametric monotone regression problem

$$Y_i = f(X_i) + \varepsilon_i, \ i = 1, \ldots, n,$$

where $f$ is monotone increasing on $[0, 1]$, and $\varepsilon_i$ are independent and identically distributed (i.i.d.) with mean 0 and finite variance $\sigma^2$, and also independent of the predictors. As a working model, we shall assume Gaussian errors to construct the likelihood, but the asymptotic properties are studied under only the sub-Gaussianity assumption. We address posterior convergence properties of a regression quantile $\mu = f^{-1}(y_0)$, where $y_0$ is a value in the interior of the range of $f$. Tang et al. [41] described the relevance of a monotone regression quantile in an engineering application involving a first-in-first-out queue. We study the contraction rate of the posterior distribution and the coverage of a credible interval obtained from the quantiles of the “projection-posterior” distribution. In this approach, the parameter space is expanded ignoring a shape or functional restriction so that a conjugate-type prior may be put, and the posterior is computationally and mathematically easily tractable. To comply with the model restriction, samples drawn from the unrestricted posterior distribution are projected on the original parameter space, which is the space of monotone functions in this context. The induced distribution is used to make an inference. A mathematical characterization of the projection procedure through the ‘switch relation’ (Groeneboom and Jongbloed [22]) is instrumental for the study of the coverage. The projection-posterior approach was also used by Lin and Dunson [32], Chakraborty and Ghosal [12, 13] for shape-restricted Bayesian inference. We establish the optimal rate of contraction $n^{-1/3}$ for $\mu$ around its true value $\mu_0 = f_0^{-1}(y_0)$ assuming that $\mu_0$ is unique, where $f_0$ is the true regression function and the condition $f_0'(\mu_0) > 0$ holds. We then find that a projection-posterior credible interval for $\mu$ has limiting coverage given by the distribution of a functional $Z_B$ of two independent two-sided standard Brownian motions $W_1$ and $W_2$, defined by

$$Z_B = \mathbb{P}(\arg\min\{W_1(t) + W_2(t) + t^2 : t \in \mathbb{R}\} \leq 0 | W_1).$$

The limit distribution is free of nuisance parameters, and hence can be computed, at least by simulation. This distribution was named as the Bayes-Chernoff distribution and its quantiles were computed in Chakraborty and
Ghosal [11] numerically. As in the case of the value of the regression function $f(x_0)$ in Chakraborty and Ghosal [11], the reverse Cox phenomenon holds, that is, the limiting coverage is higher than the credibility. Moreover, a targeted coverage $(1-\alpha)$ may be obtained by starting from a lower credibility level, which can be precisely calculated from the table of the quantiles of the Bayes-Chernoff distribution.

Further, we develop a method for more accurate Bayesian estimation of a monotone regression quantile by sampling in two stages. A fraction of the sampling budget is first used to get observations and a posterior credible interval for $\mu$ using the approach discussed above. Next, the remaining sampling budget is used to sample points only from the credible interval obtained in the first stage. Such a process of combining the information from the later set of observations with the prior knowledge from from the first stage can be adopted naturally in the Bayesian approach. By representing $f$ approximately as a linear function on this small interval and by using a Gaussian prior on the slope and the intercept of this linear function, we show that such an “active learning mechanism” can improve the posterior contraction rate of $\mu$ from $n^{-1/3}$ to $n^{-1/2}$. Such two-stage procedures in the frequentist paradigm were earlier employed by Lan et al. [30], Tang et al. [41] and Belitser et al. [6] respectively for estimating the change point of a regression function, a monotone regression quantile, and the regression mode and maximum for a smooth multivariate regression function. For the smooth regression mode problem, the two-stage rate already matches the optimal sequential rate established by Kiefer and Wolfowitz [28] for a Robbins-Monro procedure, so that expensive fully sequential sampling is unnecessary. The same two-stage rates for the regression mode and maximum were obtained for a Bayesian procedure by Yoo and Ghosal [44]. A unified setup for two-stage methods in the frequentist setting has been recently considered by Mallik et al. [33].

The rest of the paper is organized as follows. Section 2 introduces some notations, assumptions and the prior on $f$ and $\sigma$. Results on coverage of credible intervals are presented in Section 3. The two-stage Bayesian procedure is described and its contraction rate is presented in Section 4. A simulation experiment discussed in Section 5. Proofs of the results are provided in Section 6 and some auxiliary results are given in the Appendix.

2. Assumptions, prior and projection-posterior. We assume that $y_0$ belongs to the interior of the range of the true regression function $f_0$. Note that the regression quantile $\mu_0 = f_0^{-1}(y_0) \in (0, 1)$. 
Regression function. We assume that $f_0$ belong to the class $\mathcal{F}$ of monotone increasing functions on $[0, 1]$ satisfying the conditions

(i) $f'(x) > 0$ for all $x$ in a neighborhood of $\mu_0$.
(ii) for some $L > 0$ and $\kappa > 1/2$, $|f'(x_1) - f'(x_2)| \leq L|x_1 - x_2|^{\kappa}$, for all $x_1, x_2$ in a neighborhood of $\mu_0$.

The condition clearly ensures that $\mu_0$ is uniquely defined and $f_0$ is strictly increasing in a neighborhood of $\mu_0$.

Error distribution. We assume that the true distribution $P_0$ for $Y_i = f_0(X_i) + \varepsilon_i$, is given by $\varepsilon_i$ are i.i.d. sub-Gaussian with mean 0 and variance $\sigma_0^2$, and $X_i$ are i.i.d. with a bounded and positive density $g$ for $i = 1, \ldots, n$.

Prior. We use the same piecewise constant prior on $f$ Chakraborty and Ghosal [11] used, by representing $f$ as a piece-wise constant function on $I_j = ((j - 1)/J, j/J]$, $j = 1, \ldots, J$, with the number of pieces $J$ chosen appropriately depending on $n$, that is, $f = \sum_{j=1}^J \theta_j \mathbb{1}_{I_j}$, $J = J_n$. We represent the model in (1.1) as $Y = B\theta + \varepsilon$, where $Y = (Y_1, \ldots, Y_n)^T$, $B = (\{1 \{X_i \in I_j\}\})$, $\theta = (\theta_1, \ldots, \theta_J)^T$ and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$. We put the prior $\theta|\sigma \sim N_J(\zeta, \sigma^2\Lambda)$ where $\|\zeta\|_\infty$ is bounded, and $\Lambda$ is a $J \times J$ diagonal matrix with diagonal entries $\lambda_1^2, \ldots, \lambda_J^2$, and that $B_1 \leq \lambda_j \leq B_2$ for some $B_1, B_2 > 0$.

Posterior distribution. Let $N_j = \sum_{i=1}^n \mathbb{1}_{\{X_i \in I_j\}}$ stand for the counts of observations falling in the interval $I_j$, and $\bar{Y}_j = N_j^{-1} \sum_{i: X_i \in I_j} Y_i$ be the sample block-averages, $j = 1, \ldots, J$. Then the posterior distribution of $(\theta_1, \ldots, \theta_J)$ given a fixed value of $\sigma$ is the product of independent normal $\theta_j|\sigma \sim N((N_j \bar{Y}_j + \zeta_j)/(N_j + \lambda_j^{-2}), \sigma^2/(N_j + \lambda_j^{-2}))$.

Error variance. Observe that $Y|\sigma \sim N_n(B\zeta, \sigma^2(B\Lambda B^T + I_n))$, where $I_n$ is the identity matrix of order $n$. The error variance $\sigma^2$ may be estimated by maximizing the marginal likelihood of $\sigma$ or by endowing $\sigma^2$ with an inverse-gamma prior $IG(\tau_1, \tau_2)$ with parameters $\tau_1 > 2$ and $\tau_2 > 0$.

Maximizing the corresponding log-likelihood with respect to $\sigma^2$, we obtain

\begin{equation}
\hat{\sigma}_n^2 = n^{-1}(Y - B\zeta)^T(B\Lambda B^T + I_n)^{-1}(Y - B\zeta).
\end{equation}

As shown by Chakraborty and Ghosal [11], $\hat{\sigma}_n^2$ is a consistent estimator for $\sigma^2$, and the posterior distribution $IG(\tau_1 + n/2, \tau_2 + n\hat{\sigma}_n^2/2)$ of $\sigma^2$ is consistent if the inverse-gamma prior is put on it. The plug-in posterior distribution of $f$ is obtained by substituting $\hat{\sigma}_n$ for $\sigma$. The full posterior for $f$ has the identical asymptotic behavior, so we shall work only with the plug-in posterior.

Projection. For $f = \sum_{j=1}^J \theta_j \mathbb{1}_{I_j}$, let $f^* = \sum_{j=1}^J \theta^*_j \mathbb{1}_{I_j}$, where $(\theta^*_1, \ldots, \theta^*_J)$ is obtained as the minimizer of $\sum_{j=1}^J (\theta_j - \eta_j)^2$ over $\eta = (\eta_1, \ldots, \eta_J) \in \mathbb{R}^J$ subject to $\eta_1 \leq \cdots \leq \eta_J$. In other words, $f^*$ is the projection of a posterior
sample of the regression function \( f \) on the set of monotone functions. The distribution of \( f^* \) is termed as the “projection-posterior”.

The same class of functions \( f = \sum_{j=1}^{J} \theta_j 1_{I_j} \) may be used for the maximum likelihood approach, giving the sieve-maximum likelihood estimator (sieve-MLE) \( \hat{f} \) characterized as the projection of the piecewise constant function \( n^{-1} \sum_{j=1}^{J} N_j \bar{Y}_j 1_{I_j} \).

Regression quantile. For a monotone increasing function \( F \) on \([0, 1]\), we define the left-continuous inverse function \( F^{-1} : \mathbb{R} \to [0, 1] \) to be
\[
F^{-1}(y) = \inf \{ x : F(x) \geq y \}.
\]
(2.2)

Let \( h = f^*-1 \), and \( h_0 = f_0^{-1} \), be the regression quantile function and its true value, respectively. Then the object of interest is \( \mu = h(y_0) \) and the true value is \( \mu_0 = h_0(y_0) \). The goal is then to study the concentration of the induced posterior distribution of \( \mu \) at \( \mu_0 \) and the probability that a posterior quantile interval for \( \mu \) contains \( \mu_0 \).

In the maximum likelihood approach, the sieve-MLE for \( \mu \) is given by \( \hat{\mu} = \hat{f}^{-1}(y_0) \).

Notation. We use the following notations and symbols throughout the paper. For sequences \( a_n \) and \( b_n \), \( a_n \ll b_n \) means that \( a_n/b_n \to 0 \). Weak convergence or convergence in distribution for random variables is denoted by \( \Rightarrow \). Convergence in probability under a probability measure \( P \) is denoted by \( \to P \).

3. Credible set for \( \mu \) and its coverage. Before establishing the results about the coverage of a credible interval for \( \mu \), first note that the maximum marginal likelihood estimator for \( \sigma^2 \) in the plug-in Bayes approach or the marginal posterior distribution of \( \sigma^2 \) in the fully Bayes approach, are consistent uniformly for \( f_0 \in \mathcal{F} \), in view of Lemma A.1 of Chakraborty and Ghosal [11].

Let \( n^{1/3} \ll J \ll n^{2/3} \). Let \( W_1 \) and \( W_2 \) be two independent standard Brownian motions on \( \mathbb{R} \) starting at zero. Let \( Z = \arg \min \{ W_1(t) + t^2 : t \in \mathbb{R} \} \), \( \Delta_{W_1,W_2} = \arg \min \{ W_1(t) + W_2(t) + t^2 : t \in \mathbb{R} \} \), and \( C_0 = (a/b)^{2/3} \), where \( a = \sqrt{\sigma_0^2/g(\mu_0)} \) and \( b = f_0'(\mu_0)/2 \). The following theorem describes a weak limit of the posterior distribution of \( \mu \), which is a key step towards establishing the frequentist coverage of a credible interval for \( \mu \). A limit distribution with the same normalization is also obtained for the sieve-MLE.

**Theorem 3.1.** For every \( x \in \mathbb{R} \), we have \( \Pi(n^{1/3}(\mu - \mu_0) \leq x \mid D_n) \Rightarrow P(C_0 \Delta_{W_1,W_2} \leq x \mid W_1) \) and \( P_0(n^{1/3}(\hat{\mu} - \mu_0) \leq x) \Rightarrow P(C_0 Z \leq x) \).
Theorem 3.1 gives us a method of constructing a point-wise credible interval for the level set at $y_0$. For every $n \geq 1$, $\gamma \in [0, 1]$, define the $(1 - \gamma)$-posterior quantile by $Q_{n, \gamma} = \inf \{ z \in \mathbb{R} : \Pi(h(y_0) \leq z|D_n) \geq 1 - \gamma \}$. Then a two-sided $(1 - \gamma)$-posterior credible interval is given by $I_{n, \gamma} = [Q_{n, 1 - \gamma/2}, Q_{n, \gamma/2}]$. The following result quantifies the limiting coverage.

**Theorem 3.2.** For any $\gamma \in (0, 1)$, the coverage of $I_{n, \gamma}$

$$P_0(\mu_0 \in I_{n, \gamma}) \to P(\gamma/2 \leq Z_B \leq 1 - \gamma/2),$$

where $Z_B = P(\Delta_{W_1, W_2}^* \leq 0|W_1)$.

As in the case of the function value of a regression function established by Chakraborty and Ghosal [11], we observe that the limiting coverage differs from $(1 - \gamma)$. Let $A(u) = P(Z_B \leq u)$, $u \in [0, 1]$, stand for the cumulative distribution function of $Z_B$. The coverage of $I_{n, \gamma}$ is thus $A(1 - \gamma/2) = A(1 - 2A^{-1}(\gamma/2))$, which is free of any nuisance parameters of the distribution of the observations. It was argued in Chakraborty and Ghosal [11] that $A$ is continuous, strictly increasing onto $[0, 1]$ and $A(u) = 1 - A(1 - u)$, for all $u \in [0, 1]$. Hence the limiting coverage of $I_{n, \gamma}$ is $2A(1 - \gamma/2) - 1$. By simulations, Chakraborty and Ghosal [11] also showed that $A(u) > u$ for $u > 1/2$, at least at all grid points they evaluated $A$. Hence, for a regression quantile, we observe the same “reverse Cox phenomenon” that the limiting coverage is higher than the credibility $(1 - \gamma)$. Further, any targeted limiting coverage $(1 - \alpha)$ may be obtained by setting $1 - \gamma = 1 - 2A^{-1}(\alpha/2) < 1 - \alpha$. A table is provided by Chakraborty and Ghosal [11], from which we can calculate the credibility level we need to start with to obtain a targeted limiting coverage. For instance, to obtain 95% limiting coverage, the credibility should be set to 93.2%.

Also, a consequence of Theorem 3.1 is that the length of the credible interval is of the optimal order $n^{-1/3}$ in $P_0$-probability. The conclusion will be used in the next section.

**Corollary 3.3.** For any $0 < \gamma < 1$, the limiting distribution $R_{\gamma}$ of $n^{1/3}(Q_{n, \gamma} - \mu_0)$ is given by the expression

$$R_{\gamma}(x) = P(P(C_0\Delta_{W_1, W_2}^* \leq x|W_1) \geq 1 - \gamma), \quad x \in \mathbb{R}.$$

In particular, $Q_{n, \gamma} = \mu_0 + O_{P_0}(n^{-1/3})$. Further, $n^{1/3}(Q_{n, \gamma_2} - Q_{n, \gamma_1})$ has a limit distribution not degenerate at 0 for any $\gamma_1 < \gamma_2$. 
4. Two-stage sampling and accelerated contraction rate. In this section, we show that we can accelerate the posterior contraction rate for a regression quantile $\mu = f^{-1}(y_0)$ using an appropriate two-stage sampling method. For a total sampling size of $n$, we allocate only $n_1 < n$ samples to obtain the posterior distribution following the approach in Section 3, to be called the first stage posterior distribution. The remaining $n_2 = n - n_1$ samples are then obtained by sampling points uniformly from the $(1 - \gamma)$-credible region $I_{n_1, \gamma}$ for $\mu$. By Corollary 3.3, the length is small, being of the order $n^{-1/3}$ in $P_0$-probability. As the function $f$ in the vicinity of $\mu_0$ is differentiable with a positive slope, it is approximately linear on a small interval around $\mu_0$. Heuristically, ignoring the bias issue, this allows representing the parameter $\mu$ as a linear functional of the coefficients in the approximate linear regression model, provided that the observations come from a small interval, which we can ensure by sampling only from $I_{n_1, \gamma}$. Being in a parametric model approximately, the parametric rate $n^{-1/2}$, which is a dramatic improvement over the rate $n^{-1/3}$, should prevail. This motivates the two-stage sampling plan, where after the $n_1$-observations are drawn in the first stage and $I_{n_1, \gamma}$ is obtained, the remaining sampling budget is spent on drawing $n_2 := n - n_1$ observations uniformly from $I_{n_1, \gamma}$ and observing the corresponding response variables. We show that the heuristic works, in that the second stage posterior for $\mu$ contracts at the improved rate $n^{-1/2}$, provided that $f_0$ is differentiable a neighborhood of $\mu_0$, the derivative is Lipschitz continuous of order $\kappa \geq 1/2$ and that $n_1/n$ (and hence also $n_2/n$) converges to a limit strictly in $(0, 1)$. If $f_0$ is twice continuously differentiable in a neighborhood of $\mu_0$, then clearly the condition holds with $\kappa = 1$.

The conclusion above is the Bayesian counterpart of a similar frequentist result by Tang et al. [41], who sampled the second stage observations from an interval centered around a first estimate $\hat{\mu}$ based on the maximum likelihood estimator of the monotone regression function, but there the length of the interval is chosen somewhat in an ad hoc fashion. The conclusion is also similar to that in Yoo and Ghosal [44], who considered a smooth multivariate regression problem and obtained an improved posterior contraction rate for the regression mode and parametric rate for the maximum. However, a major difference is that in the smoothing problem, the second stage samples are drawn from a symmetric region around a Bayes estimator obtained in the first stage whose size needs to be chosen based on asymptotics to overcome the bias. In the present context, using a credible region to draw samples in the second stage is more natural from the Bayesian point of view. Further, it is not necessary to sample observations randomly and uniformly. For instance, observing response values corresponding to points on a discrete grid.
or randomly from another non-degenerate distribution in the interval will also suffice, with a minor adjustment of the proof.

To describe the two-stage sampling procedure, fix $p \in (0, 1)$. In the first stage, we draw $n_1$ samples $(\tilde{X}_i, \tilde{Y}_i), i = 1, \ldots, n_1$, from the model (1.1) with $\tilde{X}_1, \ldots, \tilde{X}_{n_1}$ sampled from a positive and continuous density $g_1$ on $[0, 1]$, where $n_1/n \to p$ as $n \to \infty$. We denote the first-stage data by $D_{1,n}$. We fix $0 < \gamma < 1$, and construct a $(1 - \gamma)$-credible interval $C_n$ based on the projection-posterior distribution using its $\gamma/2$ and $(1 - \gamma/2)$ quantiles as described in Section 3. Let $\tilde{\mu}$ stand for the midpoint of the credible interval and let $2\delta_n$ denote its length so that we can write $C_n = (\tilde{\mu} - \delta_n, \tilde{\mu} + \delta_n)$. We then get $n_2 = n - n_1$ independent samples $X_1, \ldots, X_{n_2}$ uniformly from $C_n$ and observe the corresponding responses $Y_1, \ldots, Y_{n_2}$ from the model (1.1).

We denote the second stage samples by $D_{2,n} = \{(X_i, Y_i) : i = 1, \ldots, n_2\}$. For the results below, it is not essential that the second stage samples are uniformly sampled, for any fixed positive and continuous density on $[0, 1]$ may be taken (for instance $g_1$, the first stage sampling density), and restricted to $C_n$.

By Corollary 3.3, $\delta_n$ is of the order $n^{-1/3}$ in $P_\theta$-probability and that $C_n$ lies inside a neighborhood of $\mu_0$ of size $n^{-1/3}$ in $P_\theta$-probability. Therefore, as the true regression function $f_0$ is differentiable at $\mu_0$ with a positive derivative, $f_0$ is approximately linear on $C_n$. This allows the nonparametric monotone regression model to be replaced by a linear model, which can be the basis of an inference on the parameter $\mu$ using the second stage samples. To analyze the second-stage posterior distribution, it is convenient to center the second stage design points at 0 by setting $Z_i = X_i - \tilde{\mu}$, $i = 1, \ldots, n_2$, so that $Z_i$, $i = 1, \ldots, n_2$, are i.i.d. uniform samples from the interval $Q_n = (-\delta_n, \delta_n)$. Note that the centered design points $Z_i$ are independent of the errors $\varepsilon$. We represent $f$ on $Q_n$ by $f(z) = \alpha + \beta z$ for $z \in Q_n$. Let $Z$ denote the $n_2 \times 2$ matrix whose $i$th row is $(1, Z_i), i = 1, \ldots, n_2$.

We use the prior $\alpha|\sigma^2 \sim N(\xi_1, \sigma^2 \nu_1)$, and independently, $\beta|\sigma \sim N(\xi_2, \sigma^2 \nu_2)$, where these parameters do not depend on $n$. Let $\xi = (\xi_1, \xi_2)^T$ and $V = \text{diag}(\nu_1, \nu_2)$. As before, we may either plug-in an estimate for $\sigma$, or assign a further prior on it.

Let $Y = (Y_1, \ldots, Y_{n_2})$ be the vector of second stage response variables. The posterior $\Pi(\alpha, \beta|D_{2,n}, \sigma^2)$ is then

$$
N_2 \left[ (Z^T Z + V^{-1})^{-1}(Z^T Y + V^{-1} \xi), \sigma^2(Z^T Z + V^{-1})^{-1} \right].
$$

The plug-in posterior of $\sigma^2$ is obtained by replacing $\sigma^2$ with its estimate $\tilde{\sigma}^2 = (n_1 \tilde{\sigma}_1^2 + n_2 \tilde{\sigma}^2)/n$ based on the aggregated data $\{D_{1,n}, D_{2,n}\}$, where $\tilde{\sigma}_1^2$ is the marginal maximum likelihood estimator for $\sigma^2$ based on the first stage
samples and \( \tilde{\sigma}^2 = \frac{1}{n_2} (Y - Z\xi)^T (ZVZ^T + I_{n_2})^{-1} (Y - Z\xi) \) is the same estimate based on the second stage samples. For the fully Bayes approach, we use the first stage posterior of \( \sigma^2 \) as a prior for the second stage. We consider \( \text{IG}(\tau_1, \tau_2) \) as the prior on \( \sigma^2 \) with \( \tau_1 > 2, \tau_2 > 0 \). We then use the resulting posterior \( \text{IG}(\tau_1 + n_1/2, \tau_2 + n\tilde{\sigma}^2_1/2) \) as prior for the second stage, yielding \( \text{IG}(\tau_1 + n/2, \tau_2 + n\tilde{\sigma}^2_2/2) \) as the second stage posterior for \( \sigma^2 \). The following proposition shows that the second stage plug-in Bayes estimator and the fully Bayes posterior of \( \sigma^2 \) are consistent, and the rates are also obtained.

**Proposition 4.1.** For \( f_0 \in \mathcal{F} \),

(a) the second stage plug-in Bayes estimator \( \tilde{\sigma}^2 \) converges to \( \sigma^2_0 \) at the rate \( n^{-1/2} \) in \( P_0 \)-probability;

(b) the second stage posterior of \( \sigma^2 \) contracts to \( \sigma^2_0 \) at the rate \( n^{-1/2} \) in \( P_0 \)-probability.

Clearly, the \( y_0 \)-regression quantile is explicitly solved in the linear model as \( \mu = \tilde{\mu} + (y_0 - \alpha)/\beta \), and hence induced posterior distribution can be used to make inference about \( \mu \). The following result establishes the parametric posterior contraction rate for \( \mu \), by using either the plug-in or the full Bayesian approach for \( \sigma^2 \).

**Theorem 4.2.** Assume that \( f_0 \) is differentiable and \( f_0' \) is Lischitz continuous of order \( \kappa \geq 1/2 \) in a neighborhood of \( \mu_0 \). Then for \( M_n \to \infty \),

\[
E_0 \Pi(|\mu - \mu_0| > M_n n^{-1/2} | D_{2,n}) \to 0.
\]

We note that the value of the credibility level \( 1 - \gamma \) plays no role in the conclusion, and in fact, the interval from where the second stage samples are drawn need not even be a fixed credible interval. It can be easily concluded from the proof that we only require that the mid-point \( \tilde{\mu} \) of the sampling interval is within \( O_{P_0}(n^{-1/3}) \) and the length \( 2\delta_n \) of the sampling interval has order between \( n^{-1/3} \) and \( n^{-1/(2(1+\kappa))} \) in \( P_0 \)-probability. For instance, \( \tilde{\mu} \) can be the MLE or the sieve MLE of \( \mu \), and \( \delta_n \) may be chosen deterministically following the above requirement.

5. Simulation results.

5.1. Coverage of the first-stage credible interval. We first study the behavior of credible intervals for \( \mu = f^{-1}(y_0) \) based on the projection-posterior \( f^* \). We first compare finite sample performance of Bayesian credible intervals with confidence intervals in terms of their coverages of \( \mu_0 = f_0^{-1}(y_0) \),
using $f_0(x) = x^2 + x/5$ for $x \in [0, 1]$, $y_0 = 0.15$, $\mu_0 = 0.3$, $G$ the uniform distribution on $[0, 1]$, and $\sigma_0 = 0.1$. We take $J$ to be the greatest integer less than or equal to $n^{1/3}\log n$. We vary sample size $n$ across four different values. For each $n$, we consider 1000 instances of the data $D_n$. For each instance, we generate 1000 posterior samples of $\theta$ and obtain the corresponding projection $f^*$ from $\theta$ and compute the inverse $f^{-1}$. We then obtain the symmetric $(1 - \alpha)$-credible interval based on these sampled values and also compute the corresponding recalibrated credible interval, that is the symmetric $(1 - \gamma)$-credible interval, with $\gamma = 2A^{-1}(\alpha/2)$, which can be calculated from Table 2 of Chakraborty and Ghosal [11]. For instance, for 95% coverage, we use 93.2% credible interval. We also compute the inverse $\hat{f}_n^{-1}(y_0)$ from the sieve-MLE $\hat{f}_n$ for each replication. We then compute the $(1 - \alpha)$-confidence intervals from $\hat{f}_n^{-1}(y_0)$ based on the quantiles of Chernoff’s distribution and the estimated value of the constant $C_0$ that appears in Theorem 3.1. We denote these three intervals by $I_B(\alpha)$, $I_F^*(\alpha)$ and $I_F(\alpha)$ respectively, and check if they contain $\mu_0$. The coverage is determined by the proportion of times the credible interval contains $\mu_0$. Let $C_B(\alpha)$, $C_B^*(\alpha)$ and $C_F(\alpha)$ denote the coverages and $L_B(\alpha)$, $L_B^*(\alpha)$ and $L_F(\alpha)$ denote the lengths of $I_B(\alpha)$, $I_B^*(\alpha)$ and $I_F(\alpha)$ respectively.

5.2. Comparison of the single and two-stage estimation errors. We compare the performance of our two-stage Bayesian procedure with the single-stage Bayesian procedure using the same total number of samples. We consider the same setup of Subsection 5.1 with the true regression function $f_0(x) = x^2 + x/5$ for $x \in [0, 1]$, $\mu_0 = 0.3$, $G$ the uniform distribution on $[0, 1]$, and $\sigma_0 = 0.1$. We take $J$ to be the greatest integer less than or equal to $n^{1/3}\log n$. We use equal splitting of the sample size across the two stages.
that is \( n_1 = n_2 = n/2 \). We study the regression quantile at \( y_0 = 0.15 \). In the first stage, we observe the responses at \( n_1 \) uniformly sampled design points from \([0, 1] \). We then sample \( n_2 \) points from the symmetric 95% credible interval. We choose \( \xi_1 = \xi_2 = 0 \) and \( \nu_1 = \nu_2 = 10 \). We use the plug-in Bayes method to estimate \( \sigma^2 \). For each \( n \), we repeat the experiment 1000 times. At each instance of the data, we compute the first stage error \( |\tilde{\mu} - \mu_0| \) and the second stage error \( |(y_0 - \tilde{\alpha})/\tilde{\beta} + \tilde{\mu} - \mu_0| \), where \( \tilde{\alpha} \) and \( \tilde{\beta} \) are the second stage posterior means of \( \alpha \) and \( \beta \) respectively. We obtain the boxplots of these 1000 errors for both the procedures.

We observe that the Bayesian two-stage procedure has considerably lower absolute error than the corresponding single stage procedure, with significant improvements evident in larger samples sizes.

6. Proofs. The proofs use a crucial identity called the switch relation: if \( \Phi \) is a lower semi-continuous function on an interval \( I \) and if \( \Phi^* \) stands for the largest convex function dominated by \( \Phi \), then for every \( t \in I, v \in \mathbb{R} \),

\[
\{ \Phi^*(t) > v \} = \{ \arg \min \{ \Phi(s) - vs : s \in I \} < t \},
\]

where \( \Phi^* \) denotes the left-derivative of \( \Phi^* \) and ‘arg min’ is the (maximum possible) minimizer. A proof of the switch relation is given in Page 56 of Groeneboom and Jongbloed [22].

In order to prove Theorem 3.1, we shall need the following result on the asymptotic distribution of \( f^* \) in a \( n^{-1/3} \)-neighborhood of \( \mu_0 \).

**Lemma 6.1.** Let \( W_1, W_2, a, b \) be as defined in Theorem 3.1. Then for \( x \in \mathbb{R} \) and \( z \in \mathbb{R} \),

\[
\Pi(n^{1/3}(f^*(\mu_0 + n^{-1/3}x) - f_0(\mu_0)) \leq z | D_n) = \sim P(\frac{a/b}{2\sqrt{3}} \arg \min_{t \in \mathbb{R}} \{W_1(t) + W_2(t) + t^2\} + z/(2b) \geq x | W_1).
\]

**Proof.** Let \( v = \mu_0 + n^{-1/3}x \). Since \( f^* \) is piece-wise constant on each \( I_j \), \( f^*(v) = \theta^*_\{v\}. \) Let \( c_n(\cdot) \) denote the graph of

\[
\left\{ (0, 0), \left( \frac{N_1}{n}, \frac{N_1}{n} \theta_1 \right), \left( \sum_{k=1}^2 \frac{N_k}{n}, \sum_{k=1}^2 \frac{N_k}{n} \theta_k \right), \ldots, \left( \sum_{k=1}^J \frac{N_k}{n}, \sum_{k=1}^J \frac{N_k}{n} \theta_k \right) \right\},
\]

with the convention that \( c_n(s) = 0 \) for \( s \leq 0 \) and \( c_n(s) = \sum_{k=1}^J (N_k/n)\theta_k \) for \( s \geq 1 \). Let \( V_n, G_n \) be stochastic processes denoting \( V_n(s) = \sum_{j=1}^{[s]} (N_j/n)\theta_j \).
Fig 1: Comparison of the mean absolute errors based on the Bayesian one-stage and two-stage procedures for $n = 50, 100, 150, 200, 250, 400, 600, 800, 1000, 1200$. 
and $G_n(s) = \sum_{j=1}^{[sJ]} N_j/n$. Now since $\theta^*_n[v,J]$ is the left-derivative of the greatest convex minorant of $c_n(\cdot)$ at the point $\sum_{k=1}^{[v,J]} N_k/n$, by the switch relation (6.1), we have that

\[
\{n^{1/3}(f^*(v) - f_0(\mu_0)) \leq z\} = \{\theta^*_n[v,J] \leq f_0(\mu_0) + n^{-1/3}z\}
\]

\[
= \{ \arg \min_{s \geq 0} \{c_n(s) - (f_0(\mu_0) + n^{-1/3}z)s \} \geq [vJ]/J \}
\]

\[
= \{ \arg \min \{V_n(j/J) - (f_0(\mu_0) + n^{-1/3}z)G_n(j/J) : j = 0, \ldots, J \} \geq [vJ] \}
\]

\[
= \{ \arg \min \{V_n(s) - (f_0(\mu_0) + n^{-1/3}z)G_n(s) : s \in [0,1] \} \geq [vJ] \};
\]

here we have used the fact that the function inside the argmin is piece-wise linear in $s$, and hence the minimum occurs at one of the points $\sum_{k=1}^{J} N_k/n = G_n(j/J), j = 1, \ldots, J$. We also note that in the last expression, $s$ can vary over the whole of $\mathbb{R}$, since $V_n(s)$ and $G_n(s)$ do not change beyond $[0,1]$.

Using the facts that $V_n(s) = G_n(s) = 0$ for $s \leq 0$, that the location of the minimum does not change upon adding a constant term or upon multiplication by a positive constant to the objective function, and applying the change of variable $s = \mu_0 + n^{-1/3}t$, \( \Pi(n^{1/3}(f^*(v) - f_0(\mu_0)) \leq z|D_n) \) can be written as

\[
\Pi(\arg \min_{t \in \mathbb{R}} \{ \frac{n^{2/3}}{g(\mu_0)} (V_n(\mu_0 + n^{-1/3}t) - V_n(\mu_0) - f_0(\mu_0)) \}
\]

\[
 = -\frac{n^{1/3}}{g(\mu_0)} z (G_n(\mu_0 + n^{-1/3}t) - G_n(\mu_0)) \geq n^{1/3} (\frac{[vJ]}{J} - \mu_0) |D_n|.
\]

We note that $n^{1/3}([vJ]/J - \mu_0) = x + O(n^{1/3}/J) \to x$, the first term inside the argmin above is

\[
\frac{n^{2/3}}{g(\mu_0)} \sum_{j=\lceil \mu_0J \rceil + 1}^{[\mu_0+1/3J],J} \frac{N_j}{n} (\theta_j - f_0(\mu_0))
\]

and the second term is given by

\[
-\frac{n^{1/3}}{g(\mu_0)} z \sum_{j=\lceil \mu_0J \rceil + 1}^{[\mu_0+1/3J],J} \frac{N_j}{n}.
\]

By Lemma 5.1 of Chakraborty and Ghosal [11], the process in (6.3) converges weakly to $(aW_1(t) + aW_2(t) + bt^2 : t \in [-K,K])$ in $L_\infty([-K,K])$, for any $K >$
0, while the process in (6.4) converges in $P_0$-probability to $-zt$, uniformly in $t \in [-K, K]$, for any given $K > 0$, by the arguments in the proof of Theorem 3.3 of Chakraborty and Ghosal [11]. Further, in view of Lemma 5.2 of Chakraborty and Ghosal [11] and the argmax theorem (Theorem 3.2.2 of van der Vaart and Wellner [42]) applied conditionally on $D_n$, we obtain that the weak limit of the expression in (6.2) is given by $P(\arg \min \{aW_1(t) + aW_2(t) + bt^2 - zt : t \in \mathbb{R}\} \geq x|W_1)$; here we have used the fact that the argmin of the process having a Chernoff distribution is a continuous variable. By Lemma A.3 of Chakraborty and Ghosal [11], this expression is the same as $P(C_0 \arg \min \{W_1(t) + W_2(t) + t^2 : t \in \mathbb{R}\} + z/(2b) \geq x|W_1)$, which proves the first assertion about the posterior distribution.

\begin{remark}
A similar conclusion is obtained for the sieve-MLE:
\begin{align*}
P_0(n^{1/3}(\hat{f}(\mu_0 + n^{-1/3}x) - f_0(\mu_0)) \leq z) \\
\to P\left(\frac{(a/b)^{2/3}}{\arg \min_{t \in \mathbb{R}}} \{W_1(t) + t^2\} + z/(2b) \geq x\right).
\end{align*}
To see this, let $\hat{c}_n(s)$ and $\hat{V}_n(s)$ be defined like $c_n(s)$ and $V_n(s)$ respectively with $\sum_{k=1}^J N_k \theta_k/n$ replaced by $\sum_{k=1}^J N_k \hat{Y}_k/n$. Then the event $\{n^{1/3}(\hat{f}(\mu_0 + n^{-1/3}x) - f_0(\mu_0)) \leq z\}$ is equivalent with the event that
\begin{align*}
\arg \min_{t \in \mathbb{R}} \left\{\frac{n^{2/3}}{g(\mu_0)}(\hat{V}_n(\mu_0 + n^{-1/3}t) - V_n(\mu_0) - f_0(\mu_0))
\right. \\
\left. - \frac{n^{1/3}}{g(\mu_0)}z(G_n(\mu_0 + n^{-1/3}t) - G_n(\mu_0))\right\} \geq n^{1/3}\left[\frac{\ceil{v/J}}{J} - \mu_0\right].
\end{align*}
The expression within the argmin is
\begin{align*}
\frac{n^{2/3}}{g(\mu_0)} \sum_{j=\lceil \mu_0J \rceil + 1}^{\lceil \mu_0 + n^{-1/3}J \rceil} \frac{N_j}{n} (\hat{Y}_j - f_0(\mu_0)) - \frac{n^{1/3}}{g(\mu_0)}z \sum_{j=\lceil \mu_0J \rceil + 1}^{\lceil \mu_0 + n^{-1/3}J \rceil} \frac{N_j}{n},
\end{align*}
which converges weakly to the process $(aW_1(t) + bt^2 - zt : t \in [-K, K])$ in $\mathbb{L}_\infty([-K, K])$ by Lemma 5.1 of Chakraborty and Ghosal [14]. In view of In view of Lemma 5.2 of Chakraborty and Ghosal [14], the argmin is tight, and hence by the argmax theorem, the weak limit of the argmax above is $\arg \min\{aW_1(t) + bt^2 - zt : t \in \mathbb{R}\}$. This is equivalent with the desired conclusion by a change of variable.

\begin{proof}[Proof of Theorem 3.1]
Since $y_0 = f_0(\mu_0)$, for any $x \in \mathbb{R}$,
\begin{align*}
\Pi(n^{1/3}(\mu - \mu_0) \leq x|D_n) = \Pi(y_0 \leq f^*(\mu_0 + n^{-1/3}x)|D_n)
\end{align*}
(6.5) \begin{align*}
= \Pi(n^{1/3}(f^*(\mu_0 + n^{-1/3}x) - f_0(\mu_0)) \geq 0|D_n),
\end{align*}
\end{proof}
which converges to $P((a/b)^{2/3} \arg \min \{W_1(t) + W_2(t) + t^2 : t \in \mathbb{R}\} \geq x|W_1)$ by Lemma 6.1.

To prove the last part of the theorem about the sieve-MLE, we use the conclusion in Remark 6.2 instead of Lemma 6.1. \hfill \square

**Proof of Theorem 3.2.** By the definition of posterior quantile $Q_{n, \gamma}$, we have that $\mu_0 \leq Q_{n, \gamma}$ if and only if $\Pi(h(y_0) \leq h_0(y_0)|D_n) \leq 1 - \gamma$. For every $x \in \mathbb{R}$, define $F_n^*(x|D_n) = \Pi(n^{1/3}(h(y_0) - h_0(y_0)) \leq x|D_n)$. Also, define $F_n^*(x|W_1) = P((a/b)^{2/3} \arg \min \{W_1(t) + W_2(t) + t^2 : t \in \mathbb{R}\} \leq x|W_1)$. Therefore,

$$
\begin{align*}
P_0(\mu_0 \leq Q_{n, \gamma}) &= P_0(\Pi(h(y_0) \leq h_0(y_0)|D_n) \leq 1 - \gamma) \\
&= P_0(F_n^*(0|D_n) \leq 1 - \gamma) \\
&\sim P(F_{a,b}^*(0|W_1) \leq 1 - \gamma).
\end{align*}
$$

The last expression can be written as $P(P(C_0 \Delta_{W_1, W_2}^{*} \leq 0|W_1) \leq 1 - \gamma) = P(Z_B \leq 1 - \gamma)$. We note that the constant $C_0$ disappears from the expression even though the limiting distribution of $(Q_{n, \gamma/2}, Q_{n, 1-\gamma/2})$ involves it, because we are evaluating the limiting distribution function at 0 and $C_0$ is a scale parameter.

For the two-sided credible interval $[Q_{n, \gamma/2}, Q_{n, 1-\gamma/2}]$, the corresponding coverage $P_0(Q_{n, \gamma/2} \leq \mu_0 \leq Q_{n, 1-\gamma/2})$ converges to $P(\gamma/2 \leq Z_B \leq 1 - \gamma/2)$ by the above arguments. \hfill \square

**Proof of Corollary 3.3.** For any $\gamma \in (0, 1)$ and $x \in \mathbb{R}$, we have that

$$
P_0(n^{1/3}(Q_{n, \gamma} - \mu_0) \leq x) = P_0(\mu_0 + n^{-1/3}x \geq Q_{n, \gamma}) \\
= P_0(\Pi(\mu \leq \mu_0 + n^{-1/3}x|D_n) \geq 1 - \gamma) \\
= P_0(\Pi(n^{-1/3}(\mu - \mu_0) \leq x|D_n) \geq 1 - \gamma) \\
\to P(P(C_0 \Delta_{W_1, W_2}^{*} \leq x|W_1) \geq 1 - \gamma),
$$

which is $R_{\gamma}(x)$. It is easily verified that $R_{\gamma}$ is a cumulative probability distribution function on $\mathbb{R}$, and hence in particular, $n^{1/3}(Q_{n, \gamma} - \mu_0)$ is tight. Thus $Q_{n, \gamma} = \mu_0 + O_{P_0}(n^{-1/3})$.

To show that $n^{1/3}(Q_{n, \gamma_2} - Q_{n, \gamma_1})$ has a limit distribution not degenerate at 0 for any pair $\gamma_1 < \gamma_2$, we first obtain the limiting joint distribution of
position 4.1 and Theorem 4.2. Recall that \( \tilde{\mu} \) of the order \( n \mu \alpha \) to be the shifted true function. Let \((V, Z)\) as 0 and 1. The centered design points

and \( \beta \alpha \) \( \mu \) \( n \nabla \) \( x \) \( x \) \( \tilde{\mu} \) is strictly increasing on \([0, 1]\) (cf., the discussion preceding Lemma 3.5 of Chakraborty and Ghosal [11]).

We present some preliminary results before giving the proofs of Proposition 4.1 and Theorem 4.2. Recall that \( \tilde{\mu} \), the midpoint of \( I_{n, \gamma} \), satisfies \( \tilde{\mu} = \mu_0 + O_{P_0}(n^{-1/3}) \) and the width 2\( \delta_n \) of \( I_{n, \gamma} := (\tilde{\mu} - \delta_n, \tilde{\mu} + \delta_n) \) is exactly of the order \( n^{-1/3} \) in probability by Corollary 3.3. Define \( f_0(\tilde{\mu} - \mu) = f_0(x) \) to be the shifted true function. Let \((\alpha_0, \beta_0)\) be a (random) vector such that \( \alpha_0 + \beta_0(x - \tilde{\mu}) = f_0(\mu_0) + f'_0(\mu_0)(x - \mu_0) \), that is, \( \alpha_0 = f_0(\mu_0) + f'_0(\mu_0)(\tilde{\mu} - \mu_0) \) and \( \beta_0 = f'_0(\mu_0) \); here \((\alpha_0, \beta_0)\) is random because the shift \( \tilde{\mu} \) depends on the first-stage samples. Let \( Z \) stand for the design matrix with columns labeled as 0 and 1. The centered design points \( Z_1, \ldots, Z_{n_2} \) are i.i.d. uniform samples and so by Lemma 6.3 below, \( Z^T Z \) is invertible with probability tending to 1 as \( n \to \infty \). Moreover, because of the regularization factor in the form of the prior covariance matrix \( V \), \( Z^T Z + V^{-1} \) is always invertible.

Let \( F_0 = (f_0(X_1), \ldots, f_0(X_{n_2}))^T \). Then by Taylor’s expansion of \( f_0 \) around \( \mu_0 \), we have that the maximum bias of the linear approximation is

\[
\left\| F_0 - Z \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \right\|_\infty = \max_{1 \leq i \leq n_2} |f_0(X_i) - f_0(\mu_0) - f'_0(\mu_0)(X_i - \mu_0)| = O_{P_0}(\delta_1^{1+\kappa})
\]

by the facts that \(|X_i - \tilde{\mu}| \leq \delta_n \) and \(|\tilde{\mu} - \mu_0| = O_{P_0}(n^{-1/3}) = O_{P_0}(\delta_n) \), and by the assumption that \( f'_0 \) is Lipschitz continuous with index \( \kappa \) in a neighborhood of \( \mu_0 \).

The following lemma describes the asymptotic behavior of the entries of \((Z^T Z)^{-1}\), and will be instrumental in bounding the second stage posterior mean and variance.
5.1 of Yoo and Ghosal [44]. The only differences from that proof are that a value of $\sigma_0$ lies between two fixed positive definite matrices in probability. The conclusion follows immediately from this. \hfill \Box

Proof of Proposition 4.1. We shall adapt the proof of Proposition 5.1 of Yoo and Ghosal [44]. The only differences from that proof are that the dimension is 1 (instead of a general $d$), the function is monotone instead of smooth, and the order of the squared bias is $\delta_n^{2(1+\kappa)}$ (instead of a power dependent on the smoothness). Using the estimates in Lemma 6.3 and the above facts, similar arguments gives the stated rate $\max(n^{-1/2},n^{-(1+\kappa)/3})$ of convergence for the second stage variance.

For the fully Bayes approach using the inverse-gamma prior, the same rate is obtained for the posterior contraction by the Markov inequality. \hfill \Box

For the remaining part of this section, we shall use $K_n$ to denote a shrinking neighborhood of $\sigma_0$ such that $\Pi(K_n|D_n) \to 0$. It suffices to condition on a value of $\sigma$ in $K_n$. We now evaluate the second stage contraction rate of $(\alpha, \beta)$ at $(\alpha_0, \beta_0)$ (conditional on the first-stage samples).

Lemma 6.4. For any $\sigma \in K_n$, if $\kappa \geq 1/2$, then

\[
\begin{align*}
E[|\alpha - \alpha_0| | D_n, \sigma] &= O_{P_0}(\max\{n^{-1/2}, \delta_n^{1+\kappa}, n^{-1}\delta_n^{-1}\}) = O_{P_0}(n^{-1/2}), \\
E[|\beta - \beta_0| | D_n, \sigma] &= O_{P_0}(\max\{n^{-1/2}\delta_n^{-1}, \delta_n^{\kappa}, n^{-1}\delta_n^{-2}\}) = O_{P_0}(n^{-1/6}).
\end{align*}
\]

Proof. By Lemma 6.3, uniformly for $\sigma \in K_n$,

\[
\begin{align*}
\text{Var}(\alpha|D_n, \sigma) &= (\sigma_0^2 + o(1))[\langle Z^T Z + V^{-1} \rangle]_{00} \lesssim [\langle Z^T Z \rangle]_{00} \lesssim n^{-1}, \\
\text{Var}(\beta|D_n, \sigma) &= (\sigma_0^2 + o(1))[\langle Z^T Z + V^{-1} \rangle]_{11} \lesssim [\langle Z^T Z \rangle]_{11} \lesssim n^{-1}\delta_n^{-2}.
\end{align*}
\]

The bias of the conditional posterior mean of $(\alpha, \beta)$ can be written as

\[
E\left[ \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \bigg| D_{2n}, \sigma \right] - \begin{pmatrix} \alpha_0 \\ \beta_0 \end{pmatrix} = (Z^T Z + V^{-1})^{-1} \left[ Z^T \varepsilon + Z^T(F_0 - Z \begin{pmatrix} \alpha_0 \\ \beta_0 \end{pmatrix}) + V^{-1}(\mathbf{x} - \begin{pmatrix} \alpha_0 \\ \beta_0 \end{pmatrix}) \right].
\]
The first term has expectation equal to zero, and the \( j \)th diagonal entry of the covariance matrix \( \sigma^2 \sum_{i=1}^{n} \left[ (Z^T Z + V^{-1})^{-1} Z^T Z (Z^T Z + V^{-1})^{-1} \right]_{jj} \lesssim n^{-1} \delta_n^{-2j}, \ j = 0, 1 \). This implies that the \( j \)th entry of the first term is \( O_P(n^{-1/2} \delta_n^{-j}), \ j = 0, 1 \).

The other two terms are deterministic. From the Cauchy-Schwarz inequality and Lemma 6.3, it is immediate that the \((0, 1)\)th entry of \((Z^T Z + V^{-1})^{-1}\) is of the order \( n^{-1} \delta_n^{-1} \). As all entries of the first column of \( Z \) are 1 and those of the second column are of the order of \( \delta_n \), it is immediate that \((Z^T Z + V^{-1})^{-1} Z^T\) is a \( 2 \times n \)-matrix with the first row having all entries bounded by a constant multiple of \( n^{-1} \), while the second row has all entries bounded by a constant multiple of \((n \delta_n)^{-1}\). Using the estimate of the bias \( \|F_0 - Z \left( \begin{array}{c} \alpha_0 \\ \beta_0 \end{array} \right) \| \), it is then easy to see that the zeroth entry of the second term is of the order of \( \delta_n^{1+\kappa} \) and the first entry is of the order of \( \delta_n^\kappa \). Both entries of the third term are bounded. Then it is very easy to conclude that the zeroth entry of the third term is at most \( n^{-1} \delta_n^{-1} \) and the first entry is at most \( n^{-1} \delta_n^{-2} \). This completes the proof.

\[ \text{PROOF OF THEOREM 4.2.} \] \( f_{\alpha_0, \beta_0}(x) = \alpha_0 + \beta_0 x \). Using the fact that \( \alpha_0 = f_0(\mu_0) + f_0'(\mu_0)(\tilde{\mu} - \mu_0) \) and \( \beta_0 = f_0'(\mu_0) \), we get

\[
\frac{f^{-1}(y_0) - f_{\alpha_0, \beta_0}^{-1}(y_0)}{\beta} = \frac{y_0 - \alpha}{\beta} - \frac{y_0 - \alpha_0}{\beta_0}
\]

\[
= \left( f_0(\mu_0) - \alpha_0 \right) \left( \frac{1}{\beta} - \frac{1}{\beta_0} \right) - \left( \alpha - \alpha_0 \right) \left( \frac{1}{\beta} - \frac{1}{\beta_0} \right) - \frac{\alpha - \alpha_0}{\beta_0}
\]

\[
= -f_0'(\mu_0)(\tilde{\mu} - \mu_0) \left( \frac{1}{\beta} - \frac{1}{\beta_0} \right) - (\alpha - \alpha_0) \left( \frac{1}{\beta} - \frac{1}{\beta_0} \right) - \frac{\alpha - \alpha_0}{f_0'(\mu_0)}
\]

Using the facts that \( |\tilde{\mu} - \mu_0| = O_P(n^{-1/3}) \), and the conclusions from Lemma 6.4 that \( |\alpha - \alpha_0| = O_P(n^{-1/2}) \) and \( |\beta - \beta_0| = O_P(n^{-1/6}) \), the desired conclusion follows.

\[ \text{References.} \]


