

# Supplement to Bayesian mode and maximum estimation and accelerated rates of contraction

WILLIAM WEIMIN YOO\*<sup>1</sup>

<sup>1</sup>*Mathematical Institute, Leiden University, P.O. Box 9512, 2300 RA Leiden, The Netherlands*  
*E-mail: [yooweimin0203@gmail.com](mailto:yooweimin0203@gmail.com)*  
 and

SUBHASHIS GHOSAL<sup>2</sup>

<sup>2</sup>*Department of Statistics, North Carolina State University, 4276 SAS Hall, 2311 Stinson Drive, Raleigh, North Carolina 27695-8203, USA*  
*E-mail: [sghosal@ncsu.edu](mailto:sghosal@ncsu.edu)*

The supplementary file contains detailed proofs of Corollary 4.2, Proposition 5.1 and Corollary 8.4. in the main paper Yoo and Ghosal [4].

**Proof of Corollary 4.2.** From the proof of Theorem 4.1 before, we know that  $\boldsymbol{\mu} - \boldsymbol{\mu}_0 = \mathbf{H}f_0(\boldsymbol{\mu}^*)^{-1}(\nabla f_0(\boldsymbol{\mu}) - \nabla f_0(\boldsymbol{\mu}_0))$ . Noting that  $\nabla f_0(\boldsymbol{\mu}_0) = \nabla f(\boldsymbol{\mu}) = \mathbf{0}$  by Assumption 2, we can use the fact  $\|\mathbf{A}\mathbf{b}\|^2 \geq \lambda_{\min}(\mathbf{A}^T\mathbf{A})\|\mathbf{b}\|^2$  to write

$$\begin{aligned} \|\boldsymbol{\mu} - \boldsymbol{\mu}_0\| &\geq \sqrt{\lambda_{\max}^{-2}\{\mathbf{H}f_0(\boldsymbol{\mu}^*)\}}\|\nabla f_0(\boldsymbol{\mu}) - \nabla f_0(\boldsymbol{\mu}_0)\| \\ &\geq \lambda_1^{-1}\|\nabla f_0(\boldsymbol{\mu}) - \nabla f(\boldsymbol{\mu})\|, \end{aligned}$$

by posterior consistency of  $\boldsymbol{\mu}^*$  as established in the proof of Theorem 5.2. Let  $\delta_n \rightarrow 0$  be some sequence. Then for some small enough constant  $h > 0$  to be determined below, we have

$$\begin{aligned} \Pi(\|\boldsymbol{\mu} - \boldsymbol{\mu}_0\| \leq h\epsilon_n | \mathbf{Y}) &\leq \Pi(\|\nabla f_0(\boldsymbol{\mu}) - \nabla f(\boldsymbol{\mu})\| \leq \lambda_1 h\epsilon_n, \|\boldsymbol{\mu} - \boldsymbol{\mu}_0\| \leq \delta_n | \mathbf{Y}) \\ &\quad + \Pi(\|\boldsymbol{\mu} - \boldsymbol{\mu}_0\| > \delta_n | \mathbf{Y}). \end{aligned}$$

Since the posterior of  $\boldsymbol{\mu}$  is consistent, the second term is  $o_{P_0}(1)$ . Using the definition of continuity of  $\mathbf{x} \mapsto \|\nabla f_0(\mathbf{x}) - \nabla f(\mathbf{x})\|$  at  $\boldsymbol{\mu}_0$  and by taking  $n$  large enough (so that  $\delta_n$  is small enough), we see that

$$\Pi(\|\boldsymbol{\mu} - \boldsymbol{\mu}_0\| \leq h\epsilon_n | \mathbf{Y}) \leq \Pi(\|\nabla f_0(\boldsymbol{\mu}_0) - \nabla f(\boldsymbol{\mu}_0)\| \leq 2\lambda_1 h\epsilon_n | \mathbf{Y}) + o_{P_0}(1).$$

To obtain the same rate as the upper bound presented in (4.3) of Theorem 4.1, we then need the lower bound point-wise version of Theorem 9.1, namely for some constant

$m_0 > 0$  and for any  $\mathbf{x} \in [0, 1]^d$ ,

$$\sup_{\|f_0\|_{\alpha, \infty} \leq R} \mathbb{E}_0 \Pi \left( |D^{\mathbf{r}} f(\mathbf{x}) - D^{\mathbf{r}} f_0(\mathbf{x})| \leq m_0 n^{-\alpha^* \{1 - \sum_{k=1}^d (r_k / \alpha_k)\} / (2\alpha^* + d)} \middle| \mathbf{Y} \right) \rightarrow 0. \quad (1)$$

One can proceed to establish such lower bound directly since we have analytical expression for the Gaussian posterior distribution. By taking  $\mathbf{r} = \mathbf{e}_k$  and  $h \leq m_0 / (2\lambda_1)$ , we conclude that  $\epsilon_n^2 = \sum_{k=1}^d n^{-2\alpha^* (1 - \alpha_k^{-1}) / (2\alpha^* + d)} \geq \max_{1 \leq k \leq d} n^{-2\alpha^* (1 - \alpha_k^{-1}) / (2\alpha^* + d)}$ . As a result, if one adds an extra lower bound assumption (4.5), we have the lower bound:

$$\mathbb{E}_0 \Pi \left( \|\boldsymbol{\mu} - \boldsymbol{\mu}_0\| \geq h n^{-\alpha^* \{1 - (\min_{1 \leq k \leq d} \alpha_k)^{-1}\} / (2\alpha^* + d)} \middle| \mathbf{Y} \right) \rightarrow 1,$$

for a small enough constant  $h > 0$ . For the posterior lower bound of  $M$ , let  $\boldsymbol{\mu}^*$  be some point in between  $\boldsymbol{\mu}$  and  $\boldsymbol{\mu}_0$ . We Taylor expand  $f_0$  around  $\boldsymbol{\mu}_0$ , add and subtract  $M$  and use the reverse triangle inequality to write

$$\begin{aligned} |M_0 - M| &\geq |f_0(\boldsymbol{\mu}) - f(\boldsymbol{\mu})| + 0.5(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^T \mathbf{H} f_0(\boldsymbol{\mu}^*)(\boldsymbol{\mu} - \boldsymbol{\mu}_0) \\ &\geq |f_0(\boldsymbol{\mu}) - f(\boldsymbol{\mu})| - 0.5\lambda_1 \|\boldsymbol{\mu} - \boldsymbol{\mu}_0\|^2, \end{aligned}$$

by the extra assumption and posterior consistency of  $\boldsymbol{\mu}^*$ . Choose  $m_n = \sqrt{\log \log n}$  and define the set  $\mathcal{T} := \{\|\boldsymbol{\mu} - \boldsymbol{\mu}_0\| \leq m_n \epsilon_n\}$ . Then for  $\omega_n := n^{-\alpha^* / (2\alpha^* + d)}$  and a small enough constant  $h > 0$  to be determined below,

$$\begin{aligned} \Pi(|M_0 - M| \leq h\omega_n | \mathbf{Y}) &\leq \Pi(|f_0(\boldsymbol{\mu}) - f(\boldsymbol{\mu})| - 0.5\lambda_1 \|\boldsymbol{\mu} - \boldsymbol{\mu}_0\|^2 \leq h\omega_n, \mathcal{T} | \mathbf{Y}) \\ &\quad + \Pi(\mathcal{T}^c | \mathbf{Y}) \\ &\leq \Pi(|f_0(\boldsymbol{\mu}) - f(\boldsymbol{\mu})| \leq h\omega_n + 0.5\lambda_1 m_n^2 \epsilon_n^2 | \mathbf{Y}) + o_{P_0}(1), \end{aligned}$$

where the last term follows from (4.3) of Theorem 4.1. Using the continuity argument as before for  $\mathbf{x} \mapsto |f_0(\mathbf{x}) - f(\mathbf{x})|$  and the fact that  $h\omega_n \gg \lambda_1 m_n^2 \epsilon_n^2$  when  $\min_{1 \leq k \leq d} \alpha_k > 2$ , we can further bound the right hand side above by

$$\Pi(|f_0(\boldsymbol{\mu}) - f(\boldsymbol{\mu})| \leq 2h\omega_n | \mathbf{Y}) + o_{P_0}(1),$$

for large enough  $n$ . By setting  $\mathbf{r} = \mathbf{0}$  in (1) above, we conclude that the first term is  $o_{P_0}(1)$  when  $h \leq m_0/2$  and the second posterior statement on  $M$  is established.  $\square$

**Proof of Proposition 5.1.** By the triangle inequality,  $|\tilde{\sigma}_*^2 - \sigma_0^2| \leq |\tilde{\sigma}_1^2 - \sigma_0^2| + |\tilde{\sigma}_2^2 - \sigma_0^2|$ . By (a) of Proposition 9.5, the first term is  $O_{P_0}(\max\{n^{-1/2}, n^{-2\alpha^* / (2\alpha^* + d)}\})$ . To bound the second term, let  $\mathbf{U} = (\mathbf{ZVZ}^T + \mathbf{I}_n)^{-1}$ . By equation (33) of page 355 in Searle [2],

$$\begin{aligned} |\mathbb{E}(\tilde{\sigma}_2^2 | \boldsymbol{\theta}_0) - \sigma_0^2| &= |n^{-1} \sigma_0^2 \text{tr}(\mathbf{U}) - \sigma_0^2| + n^{-1} (\mathbf{F}_0 - \mathbf{Z}\boldsymbol{\xi})^T \mathbf{U} (\mathbf{F}_0 - \mathbf{Z}\boldsymbol{\xi}) \\ &\lesssim n^{-1} [\text{tr}(\mathbf{I}_n - \mathbf{U}) + (\mathbf{F}_0 - \mathbf{Z}\boldsymbol{\theta}_0)^T \mathbf{U} (\mathbf{F}_0 - \mathbf{Z}\boldsymbol{\theta}_0) \\ &\quad + (\mathbf{Z}\boldsymbol{\theta}_0 - \mathbf{Z}\boldsymbol{\xi})^T \mathbf{U} (\mathbf{Z}\boldsymbol{\theta}_0 - \mathbf{Z}\boldsymbol{\xi})], \end{aligned} \quad (2)$$

where we have used  $(\mathbf{x} + \mathbf{y})^T \mathbf{G} (\mathbf{x} + \mathbf{y}) \leq 2\mathbf{x}^T \mathbf{G} \mathbf{x} + 2\mathbf{y}^T \mathbf{G} \mathbf{y}$  for any matrix  $\mathbf{G} \geq \mathbf{0}$  (Cauchy-Schwarz and the geometric-arithmetic inequalities). Let  $\mathbf{P}_Z = \mathbf{Z}(\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T$  be the orthogonal projection matrix. For matrices  $\mathbf{Q}, \mathbf{C}, \mathbf{T}, \mathbf{W}$ , the binomial inverse theorem (see Theorem 18.2.8 of Harville [1]) says that

$$(\mathbf{Q} + \mathbf{C}\mathbf{T}\mathbf{W})^{-1} = \mathbf{Q}^{-1} - \mathbf{Q}^{-1} \mathbf{C} (\mathbf{T}^{-1} + \mathbf{W}\mathbf{Q}^{-1} \mathbf{C})^{-1} \mathbf{W}\mathbf{Q}^{-1}.$$

Applying the above twice to  $\mathbf{U}$  yields

$$(\mathbf{Z}\mathbf{V}\mathbf{Z}^T + \mathbf{I}_n)^{-1} = \mathbf{I}_n - \mathbf{Z}(\mathbf{Z}^T \mathbf{Z} + \mathbf{V}^{-1})^{-1} \mathbf{Z}^T = \mathbf{I}_n - \mathbf{P}_Z + \mathbf{M}, \quad (3)$$

where  $\mathbf{M} = \mathbf{Z}(\mathbf{Z}^T \mathbf{Z})^{-1} [\mathbf{V} + (\mathbf{Z}^T \mathbf{Z})^{-1}]^{-1} (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \geq \mathbf{0}$ . Hence the first term in (2) is  $n^{-1} \text{tr}(\mathbf{P}_Z - \mathbf{M}) \leq n^{-1} \text{tr}(\mathbf{P}_Z) = (W+1)/n$ . Note that  $\mathbf{U} \leq \mathbf{I}_n$  since  $\mathbf{Z}\mathbf{V}\mathbf{Z}^T \geq \mathbf{0}$ , and the second term in (2) is bounded by

$$n^{-1} \|\mathbf{U}\|_{(2,2)} \|\mathbf{F}_0 - \mathbf{Z}\boldsymbol{\theta}_0\|^2 \leq \|\mathbf{F}_0 - \mathbf{Z}\boldsymbol{\theta}_0\|_\infty^2 \lesssim \sum_{k=1}^d \delta_{n,k}^{2\alpha_k},$$

in view of (8.3). By (3) and  $(\mathbf{I} - \mathbf{P}_Z)\mathbf{Z} = \mathbf{0}$ , the last term in (2) is  $n^{-1}(\boldsymbol{\theta}_0 - \boldsymbol{\xi})^T [\mathbf{V} + (\mathbf{Z}^T \mathbf{Z})^{-1}]^{-1} (\boldsymbol{\theta}_0 - \boldsymbol{\xi}) \leq n^{-1} \sum_{j=0}^W \delta_n^{i_j} (\theta_{0,i_j} - \xi_{i_j})^2 = O_{P_0}(n^{-1})$ , since  $\delta_{n,k} = o(1)$ ,  $k = 1, \dots, d$ ,  $\theta_{0,i_j} = O_{P_0}(1)$  and  $\xi_{i_j} = O(1)$  by assumption on the prior for any  $0 \leq j \leq W$ . Combining the three bounds established into (2), we obtain  $|\mathbb{E}(\tilde{\sigma}_2^2 | \boldsymbol{\theta}_0) - \sigma_0^2| \lesssim n^{-1} + \sum_{k=1}^d \delta_{n,k}^{2\alpha_k}$ .

We write  $n\tilde{\sigma}_2^2 = (\mathbf{F}_0 - \mathbf{Z}\boldsymbol{\xi})^T \mathbf{U} (\mathbf{F}_0 - \mathbf{Z}\boldsymbol{\xi}) + 2(\mathbf{F}_0 - \mathbf{Z}\boldsymbol{\xi})^T \mathbf{U} \boldsymbol{\varepsilon} + \boldsymbol{\varepsilon}^T \mathbf{U} \boldsymbol{\varepsilon}$  by substituting  $\mathbf{Y} = \mathbf{F}_0 + \boldsymbol{\varepsilon}$ . Observe that  $\boldsymbol{\varepsilon}$  and  $\boldsymbol{\theta}_0$  are independent by definition. Using the inequality  $\text{Var}(A_1 + A_2) \leq 2\text{Var}(A_1) + 2\text{Var}(A_2)$  (from Cauchy-Schwarz and geometric-arithmetic inequalities), we conclude that  $\text{Var}(\tilde{\sigma}_2^2 | \boldsymbol{\theta}_0)$  is bounded up to a constant multiple by

$$n^{-2} [(\mathbf{F}_0 - \mathbf{Z}\boldsymbol{\theta}_0)^T \mathbf{U}^2 (\mathbf{F}_0 - \mathbf{Z}\boldsymbol{\theta}_0) + (\mathbf{Z}\boldsymbol{\theta}_0 - \mathbf{Z}\boldsymbol{\xi})^T \mathbf{U}^2 (\mathbf{Z}\boldsymbol{\theta}_0 - \mathbf{Z}\boldsymbol{\xi}) + \text{Var}(\boldsymbol{\varepsilon}^T \mathbf{U} \boldsymbol{\varepsilon})]. \quad (4)$$

In view of (8.3) and  $\mathbf{U} \leq \mathbf{I}_n$ , the first term is bounded by  $n^{-2} \|\mathbf{U}\|_{(2,2)}^2 \|\mathbf{F}_0 - \mathbf{Z}\boldsymbol{\theta}_0\|^2 \leq n^{-1} \|\mathbf{F}_0 - \mathbf{Z}\boldsymbol{\theta}_0\|_\infty^2 \lesssim n^{-1} \sum_{k=1}^d \delta_{n,k}^{2\alpha_k}$ . Observe that since  $\mathbf{V} \geq \mathbf{0}$ ,

$$\begin{aligned} \mathbf{Z}^T \mathbf{M}^2 \mathbf{Z} &= [\mathbf{V} + (\mathbf{Z}^T \mathbf{Z})^{-1}]^{-1} (\mathbf{Z}^T \mathbf{Z})^{-1} [\mathbf{V} + (\mathbf{Z}^T \mathbf{Z})^{-1}]^{-1} \\ &\leq [\mathbf{V} + (\mathbf{Z}^T \mathbf{Z})^{-1}]^{-1} \leq \mathbf{Z}^T \mathbf{Z}. \end{aligned} \quad (5)$$

Using (3), idempotency of  $\mathbf{I}_n - \mathbf{P}_Z$  and  $(\mathbf{I}_n - \mathbf{P}_Z)\mathbf{Z} = \mathbf{0}$ , the second term in (4) is  $n^{-2} (\boldsymbol{\theta}_0 - \boldsymbol{\xi})^T \mathbf{Z}^T (\mathbf{I}_n - \mathbf{P}_Z + \mathbf{M})^2 \mathbf{Z} (\boldsymbol{\theta}_0 - \boldsymbol{\xi})$ , which is

$$n^{-2} (\boldsymbol{\theta}_0 - \boldsymbol{\xi})^T \mathbf{Z}^T \mathbf{M}^2 \mathbf{Z} (\boldsymbol{\theta}_0 - \boldsymbol{\xi}) \leq n^{-2} (\boldsymbol{\theta}_0 - \boldsymbol{\xi})^T \mathbf{Z}^T \mathbf{Z} (\boldsymbol{\theta}_0 - \boldsymbol{\xi}), \quad (6)$$

in view of (5). By (8.4) in the proof of Lemma 8.1, we can write  $\mathbf{Z}^T \mathbf{Z} = n_2 \boldsymbol{\Delta} \mathbf{A} \boldsymbol{\Delta}$  where  $\boldsymbol{\Delta} = \text{diag}\{\delta_n^{i_j} : j = 0, \dots, W\}$  and  $\mathbf{A} \rightarrow \text{EUU}^T$  in probability entry-wise, where  $\mathbf{U} = (\mathbf{U}^{i_0}, \dots, \mathbf{U}^{i_W})^T$  for  $\mathbf{U} = (U_1, \dots, U_d)^T \sim \text{Uniform}[-1, 1]^d$ . This gives  $\|\mathbf{A}\|_{(2,2)} \rightarrow$

$\|\mathbf{E}\mathbf{U}\mathbf{U}^T\|_{(2,2)}$  in probability. The entries of  $\mathbf{E}\mathbf{U}\mathbf{U}^T$  are mixed moments of  $\text{Uniform}[-1, 1]$  and hence the matrix is nonsingular with  $\|\mathbf{E}\mathbf{U}\mathbf{U}^T\|_{(2,2)} < \infty$ . Since  $\|\mathbf{\Delta}\|_{(2,2)} = 1$  and  $n_2 \leq n$ , the right hand side of (6) is bounded by

$$n_2 n^{-2} \|\mathbf{A}\|_{(2,2)} \|\mathbf{\Delta}\|_{(2,2)}^2 \|\boldsymbol{\theta}_0 - \boldsymbol{\xi}\|^2 = O_{P_0}(n^{-1}),$$

because  $\|\boldsymbol{\theta}_0 - \boldsymbol{\xi}\| \leq \|\boldsymbol{\theta}_0\| + \|\boldsymbol{\xi}\| = O_{P_0}(1)$ . By Lemma A.10 of Yoo and Ghosal [3] with  $\|\mathbf{U}\|_{(2,2)} \leq 1$  and Gaussian errors by Assumption 1, the last term in (4) is  $O(1/n)$ . Combining this with the three bounds established above, we obtain  $\text{Var}(\tilde{\sigma}_2^2 | \boldsymbol{\theta}_0) = O_{P_0}(1/n)$ . Therefore, the mean square error is  $\mathbf{E}_0(\tilde{\sigma}_2^2 - \sigma_0^2)^2 = \mathbf{E}\{\mathbf{E}[(\tilde{\sigma}_2^2 - \sigma_0^2)^2 | \boldsymbol{\theta}_0]\} \lesssim n^{-1} + \sum_{k=1}^d \delta_{n,k}^{4\alpha_k}$ .

To prove (b), observe that  $\mathbf{E}(\sigma^2 | \mathbf{Y}) \lesssim n^{-1} + \tilde{\sigma}_*^2$  and  $\text{Var}(\sigma^2 | \mathbf{Y}) \lesssim n^{-3} + n^{-1} \tilde{\sigma}_*^4$ . Therefore by Markov's inequality, the second stage posterior of  $\sigma^2$  concentrates around the second stage empirical Bayes estimator  $\tilde{\sigma}_*^2$ , and thus (b) will inherit the rate from (a) as established above.  $\square$

**Proof of Corollary 8.4.** By (8.7), we have

$$\begin{aligned} \|D^r f_{\boldsymbol{\theta}} - D^r f_{\boldsymbol{\theta}_0}\|_{\infty} &= \sup_{\mathbf{x} \in \mathcal{Q}} |D^r f_{\boldsymbol{\theta}}(\mathbf{x}) - D^r f_{\boldsymbol{\theta}_0}(\mathbf{x})| \\ &\lesssim |\theta_r - \theta_{0,r}| + \sum_{r \leq \mathbf{i} \leq \mathbf{m}_{\boldsymbol{\alpha}}, \mathbf{i} \neq r} |\theta_{\mathbf{i}} - \theta_{0,\mathbf{i}}| \delta_n^{i-r}. \end{aligned} \quad (7)$$

Hence, the upper bound (8.8) is applicable and uniformly over  $\|f_0\|_{\boldsymbol{\alpha}, \infty} \leq R$ , we will have  $\mathbf{E}_0 \sup_{\sigma^2 \in \mathcal{K}_n} \mathbf{E}[\|D^r f_{\boldsymbol{\theta}} - D^r f_{\boldsymbol{\theta}_0}\|_{\infty}^2 | \mathbf{Y}, \sigma^2] \lesssim \delta_n^{-2r} (n^{-1} + \sum_{k=1}^d \delta_{n,k}^{2\alpha_k})$ . Moreover, since the bound in (8.9) is uniform for all  $\mathbf{x} \in \mathcal{Q}$ , this implies that  $\mathbf{E}_0 \|D^r f_{\boldsymbol{\theta}_0} - D^r f_{0,z}\|_{\infty}^2 \lesssim \sum_{k=1}^d \delta_{n,k}^{2\alpha_k - 2r_k}$ . Therefore, we conclude that uniformly over  $\|f_0\|_{\boldsymbol{\alpha}, \infty} \leq R$ ,

$$\begin{aligned} &\mathbf{E}_0 \sup_{\sigma^2 \in \mathcal{K}_n} \mathbf{E}[\|D^r f_{\boldsymbol{\theta}} - D^r f_{0,z}\|_{\infty}^2 | \mathbf{Y}, \sigma^2] \\ &\lesssim \mathbf{E}_0 \sup_{\sigma^2 \in \mathcal{K}_n} \mathbf{E}[\|D^r f_{\boldsymbol{\theta}} - D^r f_{\boldsymbol{\theta}_0}\|_{\infty}^2 | \mathbf{Y}, \sigma^2] + \mathbf{E}_0 \|D^r f_{\boldsymbol{\theta}_0} - D^r f_{0,z}\|_{\infty}^2 \\ &\lesssim \delta_n^{-2r} \left( \frac{1}{n} + \sum_{k=1}^d \delta_{n,k}^{2\alpha_k} \right). \end{aligned}$$

The empirical and hierarchical posterior contraction rates then follow from (8.10) and (8.11) with absolute values replaced by sup-norms.  $\square$

## References

- [1] Harville, D. A. (1997). *Matrix Algebra from a Statistician's Perspective*. Springer-Verlag New York, Inc.
- [2] Searle, S. R. (1982). *Matrix Algebra Useful for Statistics*. John Wiley and Sons, Inc.

- [3] Yoo, W. W. and Ghosal, S. (2016). Supremum norm posterior contraction and credible sets for nonparametric multivariate regression. *Ann. Statist.*, 44(3):1069–1102.
- [4] Yoo, W. W. and Ghosal, S. (2018). Bayesian mode and maximum estimation and accelerated rates of contraction. To appear in *Bernoulli*.