Supplementary Material for "Doubly Decomposing Nonparametric Tensor Regression"

A. Proof of Theorem 1

Here, we describe the detail and proof of Theorem 1. At the beginning, we introduce a general theory for evaluating the convergence of a Bayesian estimator.

Preliminary, we introduce some theorems from previous studies. Let P_0 be a true distribution of X, $K(f,g)$ be the Kullback-Leibler divergence, and define $V(f,g) = \int(\log(f/g))^2 f dx$. Let d be the Hellinger distance, $N(\epsilon, \mathcal{P}, d)$ be the bracketing number, and $D(\epsilon, \mathcal{P}, d)$ be the packing number. Also we consider a reproducing kernel Hilbert space (RKHS), which is a closure of linear space spanned by a kernel function. Denote by $\mathcal{H}^{(k)}$ the RKHS on $\mathcal{X}^{(k)}$.

The following theorem provides a novel tool to evaluate the Bayesian estimator by posterior contraction.

Theorem 1 (Theorem 2.1 in (Ghosal et al., 2000))**.** *Consider a posterior distribution $\Pi_n(\cdot|D_n)$ on a set $\mathcal{P}$. Let ϵ_n be a sequence such that $\epsilon_n \to 0$ and $n\epsilon_n^2 \to \infty$. Suppose that, for a constant $C > 0$ and sets $\mathcal{P}_n \subset \mathcal{P}$, we have*

1. $\log D(\epsilon_n, \mathcal{P}_n, d) \leq n\epsilon_n^2$,
2. $\Pi_n(\mathcal{P}_n \backslash \mathcal{P}) \leq \exp(-n\epsilon_n^2(C+4))$,
3. $\Pi_n\left(p : -K(p, p_0) \leq \epsilon_n^2, V(p, p_0) \leq \epsilon_n^2\right) \geq \exp(-Cn\epsilon_n^2)$.

Then, for sufficiently large C', $E\Pi_n(P : d(P, P_0) \geq C'\epsilon_n)|D_n) \to 0$.

Based on the theorem, (van der Vaart & van Zanten, 2008) provide a more useful result for the Bayesian estimator with the GP prior. They consider the estimator for an infinite dimensional parameter with the GP prior and investigated the posterior contraction of the estimator with GP. They provide the following conditions.

Condition (A) With some Banach space $(\mathcal{B}, \|\cdot\|)$ and RKHS $(\mathcal{H}, \|\cdot\|)$,

1. $\log N(\epsilon_n, B_n, \|\cdot\|) \leq Cn\epsilon_n^2$,
2. $\Pr(W \notin B_n) \leq \exp(-Cn\epsilon_n^2)$,
3. $\Pr\left(\|W - w_0\| < 2\epsilon_n\right) \geq \exp(-Cn\epsilon_n^2)$,

where W is a random element in $\mathcal{B}$ and w_0 is a true function in support of W.

When the estimator with the GP prior satisfied the above conditions, the posterior contraction of the estimator is obtained as Theorem 2.1 in (Ghosal et al., 2000).

Based on this, we obtain the following result. Consider a set of GP $\{\{W_{m,x}^{(k)} : x \in \mathcal{X}^{(k)}\}\}_{m=1,\ldots,M,k=1,\ldots,K}$ and let $\{F_x : x \in \mathcal{X}\}$ be a stochastic process that satisfies $F_x = \sum_m \sum_r \lambda_r \prod_k W_{m,x_r^{(k)}}^{(k)}$.

Also we assume that there exists a true function f_0 which is constituted by a unique set of local functions $\{w_{m,0}^{(k)}\}_{m=1,\ldots,M,k=1,\ldots,K}$.

To describe posterior contraction, we define a contraction rate $\epsilon_n^{(k)}$. It converges to zero as $n \to \infty$. Let $\phi^{(k)}(\epsilon)$ be a concentration function such that

$$\phi^{(k)}(\epsilon) := \inf_{h \in \mathcal{H}^{(k)} : \|h - w_0\| < \epsilon} \|h\|_{\mathcal{H}^{(k)}}^2 - \log \Pr(\|W^{(k)}\| < \epsilon),$$

where $\|\cdot\|_{\mathcal{H}^{(k)}}$ is the norm induced by the inner product of RKHS. We define the contraction rate with $\phi^{(k)}(\epsilon)$. We denote a sequence $\{\epsilon_n^{(k)}\}_{n,k}$ satisfying

$$\phi^{(k)}(\epsilon_n^{(k)}) \leq n(\epsilon_n^{(k)})^2.$$

The order of $\epsilon_n^{(k)}$ depends on a choice of kernel function, where the optimal minimax rate is $\epsilon_n^{(k)} = O(n^{-\beta_k/(\beta_k + I_k)})$ (Tsybakov, 2008). In the following part, we set $\tilde{\epsilon_n}^{(k)} = \tilde{\epsilon}_n$ for every k. As $k' = \arg\max_k I_k$, the $\epsilon_n^{(k)}$ satisfies the condition about the concentration function.

We also note the relation between posterior contraction and the well-known risk bound. Suppose the posterior contraction such that $E\Pi_n(\theta : d_n^2(\theta, \theta_0) \geq C\epsilon_n^2|D_n) \to 0$ holds, where θ is a parameter, θ_0 is a true value, and d_n is a bounded metric. Then we obtain the following inequality:

$$\begin{aligned} &E\Pi_n(d_n^2(\theta, \theta_0)|D_n) \\ &\leq C\epsilon_n^2 + DE\Pi_n(\theta : d_n^2(\theta, \theta_0) \geq C\epsilon_n^2|D_n), \end{aligned}$$

where D is a bound of d_n. This leads $E\Pi_n(d_n^2(\theta,\theta_0)|D_n) = O(C\epsilon_n^2)$. In addition, if $\theta \mapsto d_n^2(\theta,\theta_0)$ is convex, the Jensen's inequality provides $d_n^2(\theta,\theta_0) \leq \Pi_n(d_n^2(\theta,\theta_0)|D_n)$. By taking the expectation, we obtain

$$Ed_n^2(\theta,\theta_0) \leq E\Pi_n(d_n^2(\theta,\theta_0)|D_n) = O(C\epsilon_n^2).$$

We start to prove Theorem 1. First, we provide a lemma for functional decomposition. When the function has a form of a k-product of local functions, we bound a distance between two functions with a k-sum of distance by the local functions.

Lemma 1. *Suppose that two functions $f, g : \times_{k=1}^K \mathcal{X}_k \to \mathbb{R}$ have a form $f = \prod_k f_k$ and $g = \prod_k g_k$ with local functions $f_k, g_k : \mathcal{X}_k \to \mathbb{R}$. Then we have a bound such that*

$$\|f - g\| \leq \sum_k \|f_k - g_k\| \max\left\{\|f_k\|, \|g_k\|\right\}.$$

Proof. We show the result based on induction. When $k = 2$, we have

$$f - g = f_1 f_2 - g_1 g_2 = f_1(f_2 - g_2) + (f_1 - g_1)g_2,$$

and

$$\|f - g\| \leq \|f_1\| \|f_2 - g_2\| + \|f_1 - g_1\| \|g_2\|.$$

Thus the result holds when $k = 2$.

Assume the result holds when $k = k'$. Let $k = k' + 1$. The difference between $k' + 1$-product functions is written as

$$\begin{aligned} &f - g \\ &= f_{k'+1}\prod_{k'} f_k - g_{k'+1}\prod_{k'} g_k \\ &= \prod_{k'} f_k(f_{k'+1} - g_{k'+1}) + (\prod_{k'} f_k - \prod_{k'} g_k)g_{k'+1}. \end{aligned}$$

From this, we obtain the bound

$$\begin{aligned} &\|f - g\| \\ &\leq \left\|\prod_{k'} f_k\right\| \|f_{k'+1} - g_{k'+1}\| + \left\|\prod_{k'} f_k - \prod_{k'} g_k\right\| \|g_{k'+1}\|. \end{aligned}$$

The distance $\|\prod_{k'} f_k - \prod_{k'} g_k\|$ is decomposed recursively by the case of $k = k'$. Then we obtain the result. □

Now we provide the proof of Theorem 1. Note that $M = M^* < \infty$ in the statement in Theorem 1. Asume that $C, C', C'', \ldots$ are some positive finite constants and they are not affected by other values.

Proof. We will show that F_x satisfies the condition (A). Firstly, we check the third condition in the theorem.

According to Lemma 1, the value $\|F_x - f_0\|$ is bounded as

$$\begin{aligned} &\|F_x - f_0\| \\ &\leq \sum_m \sum_r \left\|\prod_k W_r^{(k)} - \prod_k w_{m,0}^{(k)}\right\| \\ &\leq \sum_m \sum_r \sum_k \left\|W_m^{(k)} - w_{m,0}^{(k)}\right\| \prod_{k' \neq k} \max\left\{\left\|w_{m,0}^{(k')}\right\|, \left\|W_m^{(k')}\right\|\right\}. \end{aligned}$$

By denoting $\left\|\tilde{W}_m^{(k')}\right\| := \max\left\{\left\|w_{m,0}^{(k')}\right\|, \left\|W_m^{(k')}\right\|\right\}$, we evaluate the probability $Pr(\|F - f_0\| \leq \epsilon_n)$ as

$$\begin{aligned} &\Pr(\|F - f_0\| \leq \epsilon_n) \qquad (1) \\ &\geq \Pr\left(\sum_m \sum_r \sum_k \left\|W_m^{(k)} - w_{m,0}^{(k)}\right\| \prod_{k' \neq k} \left\|\tilde{W}_m^{(k')}\right\| \leq \epsilon_n\right) \\ &\geq \Pr\left(\sum_r \sum_k \sum_m \left\|W_m^{(k)} - w_{m,0}^{(k)}\right\| \leq \frac{1}{C_k}\epsilon_n\right), \qquad (2) \end{aligned}$$

where C_k is a positive finite constant satisfying $C_k = \max_m \prod_k \left\|\tilde{W}_m^{(k')}\right\|$.

From (van der Vaart & van Zanten, 2008), we use the following inequality for every Gaussian random element W:

$$\Pr\left(\|W - w_0\| \leq \epsilon_n\right) \geq \exp(-n\epsilon_n^2),$$

Then, by seting $\epsilon_n = \sum_{m,r,k} \epsilon_n^{(k)}$ and with some constant C, we bound (2) below as

$$\begin{aligned} &\Pr\left(\sum_r \sum_k \sum_m \left\|W_m^{(k)} - w_{m,0}^{(k)}\right\| \leq \frac{1}{C_k}\sum_{m,r,k} \epsilon_n^{(k)}\right) \\ &\geq \prod_{m,r,k} \Pr\left(\left\|W_m^{(k)} - w_{m,0}^{(k)}\right\| \leq \frac{1}{C_k}\epsilon_n^{(k)}\right) \\ &\geq \prod_m \prod_r \prod_k \exp\left(-\frac{n}{(C_k)^2}\epsilon_n^{(k),2}\right) \\ &\geq \exp\left(-n\sum_{m,r,k} \epsilon_n^{(k),2}\right). \end{aligned}$$

For the second condition, we define a subspace of the Banach space as

$$B_n^{(k)} = \epsilon_n^{(k)}\mathcal{B}_1^{(k)} + M_n^{(k)}\mathcal{H}_1^{(k)},$$

for all $k = 1, \ldots, K$. Note $\mathcal{B}_1^{(k)}$ and $\mathcal{H}_1^{(k)}$ are unit balls in $\mathcal{B}$ and $\mathcal{H}$. Also, we define B_n as

$$B_n := \left\{w : w = MR\prod_k w_k, w_k \in B_n^{(k)}, \forall k\right\}.$$

As shown in (van der Vaart & van Zanten, 2008), for every r and k,

$$Pr\left(W_m^{(k)} \notin B_n^{(k)}\right) \leq 1 - \Phi(\alpha_n^{(k)} + M_n^{(k)}),$$

where Φ is the cumulative distribution function of the standard Gaussian distribution; $\alpha_n^{(k)}$ and $M_n^{(k)}$ satisfy the following equation with a constant $C' > 0$ as

$$\alpha_n^{(k)} = \Phi^{-1}(Pr(W_m^{(k)} \in \epsilon_n \mathcal{B}_1^{(k)})) = \Phi^{-1}(\exp(-\phi_0(\epsilon_n^{(k)}))),$$
$$M_n^{(k)} = -2\Phi^{-1}(\exp(-C' n (\epsilon_n^{(k)})^2)).$$

By setting $\alpha_n^{(k)} + M_n^{(k)} \geq \frac{1}{2} M_n^{(k)}$ and using the relation $\phi_0(\epsilon) \leq n\epsilon_n^2$, we have

$$Pr\left(W_m^{(k)} \notin B_n^{(k)}\right) \leq 1 - \Phi\left(\frac{1}{2} M_n^{(k)}\right)$$
$$= \exp(-C' n (\epsilon_n^{(k)})^2).$$

This leads

$$Pr(F_x \notin B_n) \leq \prod_m \prod_k \prod_r \Pr(W_m^{(k)} \notin B_n^{(k)})$$
$$\leq \prod_m \prod_k \prod_r \exp(-C' n (\epsilon_n^{(k)})^2)$$
$$= \exp(-C' \sum_{m,r,k} (\epsilon_n^{(k)})^2).$$

Finally, we show the first condition. Let $\{h_j^{(k)}\}_{j=1}^{N^{(k)}}$ be a set of elements of $M_n^{(k)} \mathcal{H}_1^{(k)}$ for all k. Also, we set that each $h_j^{(k)}$ are $2\epsilon_n^{(k)}$ separated, thus $\epsilon_n^{(k)}$ balls with center $h_j^{(k)}$ do not have intersections. According to Section 5 in (van der Vaart & van Zanten, 2008), we have

$$1 \geq \sum_{j=1}^{N^{(k)}} \Pr(W_m^{(k)} \in h_j^{(k)} + \epsilon_n B_1^{(k)})$$
$$\geq \sum_{j=1}^{N^{(k)}} \exp(-\frac{1}{2} \|h_j^{(k)}\|_{\mathcal{H}}^2) Pr(W \in \epsilon_n^{(k)} \mathcal{B}_1^{(k)})$$
$$\geq N^{(k)} \exp\left(-\frac{1}{2}(M_n^{(k)})^2\right) \exp(-\phi_0(\epsilon_n^{(k)})).$$

Consider $2\epsilon_n^{(k)}$-nets with center $\{h_j^{(k)}\}_{j=1}^{N^{(k)}}$. The nets cover $M_n^{(k)} \mathcal{H}_1^{(k)}$, we obtain

$$N(2\epsilon_n^{(k)}, M_n^{(k)} \mathcal{H}_1^{(k)}, \|\cdot\|)$$
$$\leq N^{(k)} \leq \exp\left(\frac{1}{2}(M_n^{(k)})^2\right) \exp(\phi^{(k)}(\epsilon_n^{(k)})).$$

Because every point in $B_n^{(k)}$ is within $\epsilon_n^{(k)}$ from some point of $M_n \mathcal{H}_1^{(k)}$, we have

$$N(3\epsilon_n^{(k)}, B_n^{(k)}, \|\cdot\|) \leq N(2\epsilon_n^{(k)}, M_n^{(k)} \mathcal{H}_1^{(k)}, \|\cdot\|).$$

By Lemma 1, for every elements $w, w' \in B_n$ constructed as $w = \prod_k w^{(k)}, w^{(k)} \in B_n^{(k)}$, its distance is evaluated as

$$\|w - w'\| = \left\| \prod_k w_k - \prod_k w_k' \right\|$$
$$\leq \sum_k \left\| w^{(k)} - w'^{(k)} \right\| \prod_{k' \neq k} \left\| \tilde{w}^{(k')} \right\|$$
$$\leq \prod_{k' \neq k} C_{k'} \sum_k \left\| w^{(k)} - w'^{(k)} \right\|. \quad (3)$$

We consider a set $\{h^* : h^* = \prod_{k,j} h_j^{(k)}\}$, which are the element of B_n. According to (3), the $C\epsilon_n$-net with center $\{h^*\}$ will cover B_n, and its number is equal to $\prod_k N(\epsilon_n, B_n^{(k)}, \|\cdot\|)$. Let $C \sum_{m,r,k} \epsilon_n^{(k)} =: \epsilon_n'$ and we have

$$\log N(3\epsilon_n', B_n, \|\cdot\|)$$
$$\leq \sum_{m,r,k} \log N(3\epsilon_n^{(k)}, B_n^{(k)}, \|\cdot\|)$$
$$\leq \sum_{m,r,k} \log N(2\epsilon_n^{(k)}, M_n^{(k)} \mathcal{H}_1^{(k)}, \|\cdot\|)$$
$$\leq \sum_{m,r,k} \left(\frac{1}{2}(M_n^{(k)})^2 + \phi^{(k)}(\epsilon_n^{(k)}) \right)$$
$$\leq \sum_{m,r,k} \left(C'' n (\epsilon_n^{(k)})^2 + C''' n (\epsilon_n^{(k)})^2 \right)$$
$$\leq C'''' n \sum_{m,r,k} (\epsilon_n^{(k)})^2.$$

The last inequality is from the definition of M_n and $\phi^{(k)}(\epsilon_n)$.

We check that the conditions (A) are all satisfied, thus we obtain posterior contraction of the GP estimator with rate $\epsilon_n^{(k)}$. Also, according to the connection between posterior contraction and the risk bound, we achieve the result of Theorem 1.

□

B. Proof of Theorem 2

We define the representation for the true function. Recall the notation $f = \sum_m \bar{f}_m$. We introduce the notation for the true function f^* and the GP estimator $\hat{f}$ as follow:

$$f^* = \sum_{m=1}^{\infty} \bar{f}_m^*$$
$$\hat{f}_n = \sum_{m=1}^{M} \hat{\bar{f}}_m.$$

We decompose the above two functions as follows:

$$\begin{aligned}\|f^* - \hat{f}\|_n &= \left\| \sum_{m=1}^{\infty} \bar{f}_m^* - \sum_{m=1}^{M} \hat{\bar{f}}_m \right\|_n \\ &= \left\| \sum_{m=M+1}^{\infty} \bar{f}_m^* - \sum_{m=1}^{M} (\bar{f}_m^* - \hat{\bar{f}}_m) \right\|_n \\ &\leq \left\| \sum_{m=M+1}^{\infty} \bar{f}_m^* \right\|_n + \left\| \sum_{m=1}^{M} (\bar{f}_m^* - \hat{\bar{f}}_m) \right\|_n . \end{aligned} \tag{4}$$

Consider the first term with Assumption 1. Assumption 1 provides a following relation:

$$\sum_{m=1}^{\infty} \|\bar{f}_m\|_2 m^{\gamma} < \infty. \tag{5}$$

Then, the expectation of the first term of (4) is bounded by

$$\begin{aligned} E\left\| \sum_{m=M+1}^{\infty} \bar{f}_m^* \right\|_n &\leq C \sum_{m=M+1}^{\infty} \left\|\bar{f}_m^*\right\|_2 \\ &\leq C \frac{1}{(M+1)^{\gamma}} \sum_{m=1}^{\infty} \left\|\bar{f}_m^*\right\|_2 m^{\gamma}, \end{aligned}$$

with finite constant $C > 0$. The exchangeability of the first inequality is guaranteed by the setting of f^*. Also, the second inequality comes from (5). Then, we have that

$$\frac{1}{(M+1)^{\gamma}} \sum_{m=1}^{\infty} \left\|\bar{f}_m^*\right\|_2 m^{\gamma} = O(M^{-\gamma}).$$

About the second term of (4), we consider the estimation for f^* with finite M. As shown in the proof of Theorem 1, the estimation of $\sum_m^M \bar{f}_m$ is evaluated as

$$\begin{aligned} E\|\hat{f} - f^*\|_2 &= O\left(\sum_{m}^{M} n^{-\frac{\beta}{2\beta + \max_k I_k}} \right) \\ &= O\left(M n^{-\frac{\beta}{2\beta + \max_k I_k}} \right). \end{aligned}$$

Finally, we obtain the relation

$$E\|f^* - \hat{f}\|_n = O(M^{-\gamma}) + O\left(M n^{-\frac{\beta}{2\beta + \max_k I_k}} \right).$$

Then, we allow M to increase as n increases. Let $M \asymp n^{\zeta}$ with positive constant ζ, and simple calculation concludes that $\zeta = (\frac{\beta}{2\beta + \max_k I_k})/(1+\gamma)$ is optimal. By substituting ζ, we obtain the result.

References

Ghosal, Subhashis, Ghosh, Jayanta K, and van der Vaart, Aad W. Convergence rates of posterior distributions. *Annals of Statistics*, 28(2):500–531, 2000.

Tsybakov, Alexandre B. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008. ISBN 0387790519, 9780387790510.

van der Vaart, AW and van Zanten, H. Rates of contraction of posterior distributions based on gaussian process priors. *The Annals of Statistics*, 36(3):1435–1463, 2008.