A Supplementary: On the Equivalence of Tensor Regression and Gaussian Process

A.1 Eigenvalue problem

Let $\mathbf{K} = \tilde{\mathbf{U}}\tilde{\mathbf{U}}^\top$, take derivative over $\tilde{\mathbf{U}}$, we obtain the stationary point condition: $\mathbf{y}\mathbf{y}^\top(\mathbf{K}+\mathbf{D})^{-1}\tilde{\mathbf{U}} = \tilde{\mathbf{U}}$, Given the decomposition of $\tilde{\mathbf{U}} = \mathbf{U}_x\mathbf{\Sigma}_x\mathbf{V}_x^\top$, similar to (Lawrence, 2004), we have

$$\begin{aligned}
\mathbf{y}\mathbf{y}^\top(\mathbf{K}+\mathbf{D})^{-1}\tilde{\mathbf{U}} &= \tilde{\mathbf{U}} \\
\mathbf{y}\mathbf{y}^\top(\mathbf{K}+\mathbf{D})^{-1}\mathbf{U}_x\mathbf{\Sigma}_x\mathbf{V}_x^\top &= \mathbf{U}_x\mathbf{\Sigma}_x\mathbf{V}_x^\top \\
\mathbf{y}\mathbf{y}^\top\mathbf{U}_x(\mathbf{\Sigma}_x + \mathbf{D}\mathbf{\Sigma}_x^{-1})^{-1}\mathbf{V}_x^\top &= \mathbf{U}_x\mathbf{\Sigma}_x\mathbf{V}_x^\top \\
\mathbf{y}\mathbf{y}^\top\mathbf{U}_x &= \mathbf{U}_x(\mathbf{\Sigma}_x^2 + \mathbf{D})
\end{aligned}$$

which is a eigenvalue problem in the transformed space.

A.2 Derivatives for the Optimization

Given that $\mathbf{y} \sim N(\mathbf{0}, \mathbf{K}+\mathbf{D})$, where $\mathbf{K} = \phi(\mathbf{X}) \otimes_{m=1}^M \mathbf{K}_m \phi(\mathbf{X})^\top$.

Decompose $\mathbf{K}_m = \mathbf{U}_m\mathbf{U}_m^\top$, we have $\mathbf{K} = \phi(\mathbf{X})(\otimes_{m=1}^M \mathbf{U}_m)(\otimes_{m=1}^M \mathbf{U}_m^\top)\phi(\mathbf{X})^\top$.

Let $\tilde{\mathbf{U}} = \phi(\mathbf{X})(\otimes_{m=1}^M \mathbf{U}_m)$, we have $\mathbf{K} = \tilde{\mathbf{U}}\tilde{\mathbf{U}}^\top$

The negative log-likelihood

$$L = \frac{1}{2}\mathbf{y}^\top(\tilde{\mathbf{U}}\tilde{\mathbf{U}}^\top + \mathbf{D})^{-1}\mathbf{y} + \frac{1}{2}\log\det(\tilde{\mathbf{U}}\tilde{\mathbf{U}}^\top + \mathbf{D}) + const$$

Based on Woodbury lemma, $(\tilde{\mathbf{U}}\tilde{\mathbf{U}}^\top + \mathbf{D})^{-1} = \mathbf{D}^{-1} - \mathbf{D}^{-1}\tilde{\mathbf{U}}(\mathbf{D} + \tilde{\mathbf{U}}^\top\tilde{\mathbf{U}})^{-1}\tilde{\mathbf{U}}^\top$ as well as matrix determinant lemma $\det(\tilde{\mathbf{U}}\tilde{\mathbf{U}}^\top + \mathbf{D}) = \det(\mathbf{I} + \tilde{\mathbf{U}}^\top\mathbf{D}^{-1}\tilde{\mathbf{U}})\det(\mathbf{D}) = \det(\mathbf{D} + \tilde{\mathbf{U}}^\top\tilde{\mathbf{U}})$

Denote $\mathbf{\Sigma} = \mathbf{D} + \tilde{\mathbf{U}}^\top\tilde{\mathbf{U}}$, let $\mathbf{w} = \mathbf{\Sigma}^{-1}\tilde{\mathbf{U}}^\top\mathbf{y}$. The objective function can be rewrite as

$$L = \frac{1}{2}\mathbf{D}^{-1}\mathbf{y}^\top\mathbf{y} - \frac{1}{2}\mathbf{D}^{-1}\mathbf{y}^\top\tilde{\mathbf{U}}\mathbf{\Sigma}^{-1}\tilde{\mathbf{U}}^\top\mathbf{y} + \frac{1}{2}\log\det(\mathbf{\Sigma}) + const$$

Take derivative over $\mathbf{U}_{m(i,j)}$, we have

$$\frac{\partial L}{\partial \mathbf{U}_{m(i,j)}} = \mathrm{tr}[(\frac{\partial L}{\partial \tilde{\mathbf{U}}})^\top(\frac{\partial \tilde{\mathbf{U}}}{\partial \mathbf{U}_{m(i,j)}})], \quad \frac{\partial L}{\partial \tilde{\mathbf{U}}} = \tilde{\mathbf{U}}(\mathbf{\Sigma}^{-1} + \mathbf{w}\mathbf{D}^{-1}\mathbf{w}^\top)^{-1} - \mathbf{y}\mathbf{D}^{-1}\mathbf{w}^\top$$

$$\frac{\partial \tilde{\mathbf{U}}}{\partial \mathbf{U}_{m(i,j)}} = \frac{\partial \phi(\mathbf{X})}{\partial \mathbf{U}_{m(i,j)}}(\mathbf{U}_M \otimes \cdots \frac{\partial \mathbf{U}_m}{\partial \mathbf{U}_{m(i,j)}} \cdots \otimes \mathbf{U}_1) = \frac{\partial \phi(\mathbf{X})}{\partial \mathbf{U}_{m(i,j)}}(\mathbf{U}_M \otimes \cdots \mathbf{O}_{m(i,j)} \cdots \otimes \mathbf{U}_1)$$

Here $\mathbf{O}_{m(i,j)} = \mathbf{e}_i\mathbf{e}_j^\top$ is a matrix with all zeros, but the (i,j)th entry as one.

The predictive distribution: $p(y_\star|\mathbf{x}_\star, \mathbf{X}, \mathbf{y}) \sim N(\mu_\star, \sigma_\star)$:

$$\begin{aligned}
\mu_\star &= \mathbf{k}(\mathbf{x}_\star, \mathbf{X})(\mathbf{D}^{-1} - \mathbf{D}^{-1}\tilde{\mathbf{U}}(\mathbf{D} + \tilde{\mathbf{U}}^\top\tilde{\mathbf{U}})^{-1}\tilde{\mathbf{U}}^\top)\mathbf{y} \\
\sigma_\star &= \mathbf{k}(\mathbf{x}_\star, \mathbf{x}_\star) - \mathbf{k}(\mathbf{x}_\star, \mathbf{X})(\mathbf{D}^{-1} - \mathbf{D}^{-1}\tilde{\mathbf{U}}(\mathbf{D} + \tilde{\mathbf{U}}^\top\tilde{\mathbf{U}})^{-1}\tilde{\mathbf{U}}^\top)\mathbf{k}(\mathbf{X}, \mathbf{x}_\star)
\end{aligned}$$

Where $\tilde{\mathbf{U}} = \phi(\mathbf{X})(\otimes_{m=1}^M \mathbf{U}_m)$.

A.3 Proof for Proposition 2.1

Consider a 3-mode $T_1 \times T_2 \times T_3$ tensor $\mathcal{W}$ of functions $\mathcal{W}_{(1)} = [\mathbf{w}_1(\mathbf{X}), \cdots, \mathbf{w}_T(\mathbf{X})]$

$$\mathcal{W} = \mathcal{S} \times_1 \mathbf{U}_1(\mathcal{X}) \times_2 \mathbf{U}_2 \times_3 \mathbf{U}_3$$

where $\mathbf{U}_m$ is an orthogonal $T_m \times R_m$ matrix. Assuming $\mathbf{U}_1(\mathcal{X})$ satisfies $\mathbb{E}[\mathbf{U}_1^\top \mathbf{U}_1] = \mathbf{I}$ (orthogonal design after rotation).

With Tucker property

$$\mathcal{W}_{(1)} = \mathbf{U}_1(\mathcal{X})\mathcal{S}_{(1)}(\mathbf{U}_2\mathbf{U}_3)^\top$$

The population risk can be written as

$$\mathcal{L}(\mathcal{W}) = \mathrm{tr}\Big\{(\mathcal{Y} - \langle \mathcal{X}, \mathcal{W}\rangle)(\mathcal{Y} - \langle \mathcal{X}, \mathcal{W}\rangle)^\top\Big\} = \mathrm{tr}\Big\{ \begin{pmatrix} 2\mathbf{I} \\ -\mathcal{S}_{(1)}(\mathbf{U}_2\mathbf{U}_3)^\top \end{pmatrix}^\top \mathbb{E}[\mathrm{cov}(\mathcal{Y}, \mathbf{U}_1(\mathcal{X})] \begin{pmatrix} \mathbf{0} \\ -\mathcal{S}_{(1)}(\mathbf{U}_2\mathbf{U}_3)^\top \end{pmatrix} + \mathbb{E}(\mathcal{Y}\mathcal{Y}^\top)\Big\}$$

Denote $\mathbb{E}[\mathrm{cov}(\mathcal{Y}, \mathbf{U}_1(\mathcal{X})] = \mathbf{\Sigma}(\mathbf{U}_1)$, bound the difference

$$\begin{aligned}
\mathcal{L}(\mathcal{W}) - \hat{\mathcal{L}}(\mathcal{W}) &= \mathrm{tr}\Big\{ \begin{pmatrix} -2\mathbf{I} \\ \mathcal{S}_{(1)}(\mathbf{U}_2\mathbf{U}_3)^\top \end{pmatrix} (\mathbf{\Sigma}(\mathbf{U}_1) - \hat{\mathbf{\Sigma}}(\mathbf{U}_1)) \begin{pmatrix} \mathbf{0} \\ \mathcal{S}_{(1)}(\mathbf{U}_2\mathbf{U}_3)^\top \end{pmatrix} \Big\} \\
&\leq \Big\| \begin{pmatrix} -2\mathbf{I} \\ \mathcal{S}_{(1)}(\mathbf{U}_2\mathbf{U}_3)^\top \end{pmatrix} (\mathbf{\Sigma}(\mathbf{U}_1) - \hat{\mathbf{\Sigma}}(\mathbf{U}_1)) \Big\|_2 \Big\| \begin{pmatrix} \mathbf{0} \\ \mathcal{S}_{(1)}(\mathbf{U}_2\mathbf{U}_3)^\top \end{pmatrix} \Big\|_\star \\
&\leq C \max\{2, \|\mathcal{S}_{(1)}\|_\star^2\} \|\mathbf{\Sigma}(\mathbf{U}_1) - \hat{\mathbf{\Sigma}}(\mathbf{U}_1)\|_2
\end{aligned}$$

With C as a universal constant. The inequality holds with Schatten norm Hölder's inequality

$$\|AB\|_{S_1} \leq \|A\|_{S_p}\|B\|_{S_q} \quad 1/p + 1/q = 1$$

Given that $\sup_{\mathbf{U}_1} \|\mathbf{\Sigma}(\mathbf{U}_1) - \hat{\mathbf{\Sigma}}(\mathbf{U}_1)\|_2 = \mathcal{O}_P\Big(\sqrt{\frac{T_2T_3 + \log(T_1T_2T_3)}{N}}\Big)$

Denote empirical risk $\hat{\mathcal{L}} = \sum_{t=1}^T \sum_{i=1}^{n_t} \mathcal{L}(\langle \mathbf{w}_t, \mathbf{x}_{t,i}\rangle$. Let $\mathcal{W}^\star = \inf_{\mathcal{W}\in\mathcal{C}} \mathcal{L}(\mathcal{W})$. The excess risk

$$\begin{aligned}
\mathcal{L}(\hat{\mathcal{W}}) - \mathcal{L}(\mathcal{W}^\star) &= \mathcal{L}(\hat{\mathcal{W}}) - \hat{\mathcal{L}}(\hat{\mathcal{W}}) + (\hat{\mathcal{L}}(\hat{\mathcal{W}}) - \hat{\mathcal{L}}(\mathcal{W}^\star) + (\hat{\mathcal{L}}(\mathcal{W}^\star - \mathcal{L}(\mathcal{W}^\star)) \\
&\leq [\mathcal{L}(\hat{\mathcal{W}}) - \hat{\mathcal{L}}(\hat{\mathcal{W}})] - [\mathcal{L}(\mathcal{W}^\star) - \hat{\mathcal{L}}(\mathcal{W}^\star)] \\
&\leq 2\sup_{\mathcal{W}\in\mathcal{C}_N} \{\mathcal{L}(\mathcal{W}) - \hat{\mathcal{L}}(\mathcal{W})\} \\
&\leq \mathcal{O}\Big(\|\mathcal{S}_{(1)}\|_\star^2 \|\mathbf{\Sigma}(\mathbf{U}_1) - \hat{\mathbf{\Sigma}}(\mathbf{U}_1)\|_2\Big)
\end{aligned}$$

if we assume $\|\mathcal{S}_{(1)}\|_\star^2 = o(\Big(\frac{N}{T_2T_3 + \log(T_1T_2T_3)}\Big)^{1/4})$, then $\mathcal{L}(\hat{\mathcal{W}}) - \mathcal{L}(\mathcal{W}^\star) \leq o(1)$, thus we obtain the oracle inequality as stated.

A.4 Proof of Theorem 2.2

We can extend the approach of single task Gaussian process (Sollich and Halees, 2002) to our setting. We provide the derivation for the full-rank case, but similar results apply to low-rank case as well. The Bayes error for the full-rank covariance model is:

$$\hat{\epsilon} = \mathrm{tr}(\mathbf{\Lambda}'^{-1} + \mathbf{\Psi}^\top \mathbf{D}^{-1}\mathbf{\Psi})^{-1}$$

To obtain learning curve $\epsilon = \mathbb{E}_{\mathcal{D}}[\hat{\epsilon}]$, it is useful to see how the matrix $\mathcal{G} = (\mathbf{\Lambda}^{-1} + \mathbf{\Psi}^\top \mathbf{D}^{-1}\mathbf{\Psi})^{-1}$ changes with sample size. $\mathbf{\Psi}^\top\mathbf{\Psi}$ can be interpreted as the input correlation matrix.

To account for the fluctuations of the element in $\mathbf{\Psi}^\top\mathbf{\Psi}$, we introduce auxiliary offset parameters $\{v_t\}$ into the definition of $\mathcal{G}$. Define resolvent matrix

$$\mathcal{G}^{-1} = \mathbf{\Lambda}^{-1} + \mathbf{\Psi}^\top \mathbf{D}^{-1}\mathbf{\Psi} + \sum_t v_t \mathbf{P}_t$$

where $\mathbf{P}_t$ is short for $\mathbf{P}_{t_1,\cdots,t_M}$, which defines the projection of tth task to its multi-directional indexes.

Evaluating the change

$$\mathcal{G}(n+1) - \mathcal{G}(n) = [\mathcal{G}^{-1}(n) + \sigma_t^{-2}\psi_t\psi_t^\top]^{-1} - \mathcal{G}(n) = \frac{\mathcal{G}(n)\psi_t\psi_t^\top \mathcal{G}(n)}{\sigma_t^2 + \psi_t^\top \mathcal{G}(n)\psi_t}$$

where element $(\psi_t)_i = \delta_{\tau_{n+1},t}\phi_{it}(x_{n+1})$ and τ maps the global sample index to task-specific sample index. Introducing $\mathbf{G} = \mathbb{E}_{\mathcal{D}}[\mathcal{G}]$ and take expectation over numerator and denominator separately, we have

$$\frac{\partial \mathbf{G}}{\partial n_t} = -\frac{\mathbb{E}_{\mathcal{D}}[\mathcal{G}\mathbf{P}_t\mathcal{G}]}{\sigma_t^2 + \mathrm{tr}\mathbf{P}_t\mathbf{G}}$$

Since generalization error $\epsilon_t = \mathrm{tr}\mathbf{P}_t\mathbf{G}$, we have that $-\mathbb{E}_{\mathcal{D}}[\mathcal{G}\mathbf{P}_t\mathcal{G}] = \frac{\partial}{\partial v_t}\mathbb{E}_{\mathcal{D}}[\mathcal{G}] = \frac{\partial \mathbf{G}}{v_t}$. Multiplying $\mathbf{P}_s$ on both sides yields the approximation for the expected change:

$$\frac{\partial \mathbf{P}_s\mathbf{G}}{\partial n_t} = \frac{\partial \epsilon_s}{\partial n_t} = \frac{1}{\sigma_t^2 + \epsilon_t}\frac{\partial \epsilon_s}{\partial v_t}$$

Solving $\epsilon_t(N, v)$ using the methods of characteristic curves and resetting v to zero, gives the self-consistency equations:

$$\epsilon_t(N) = \mathrm{tr}\mathbf{P}_t\Big(\mathbf{\Lambda}'^{-1} + \sum_s \frac{n_s}{\sigma_s^2 + \epsilon_s}\Big)^{-1}$$