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

A.1 Eigenvalue problem

Let $K = \hat{U}\hat{U}^T$, take derivative over $\hat{U}$, we obtain the stationary point condition: $yy^T(K + D)^{-1}\hat{U} = \hat{U}$, Given the decomposition of $\hat{U} = U_x\Sigma_xV_x^T$, similar to [Lawrence 2004], we have

$$\begin{align*}
yy^T(K + D)^{-1}\hat{U} &= \hat{U} \\
\hat{U}(K + D)^{-1}U_x\Sigma_xV_x^T &= U_x\Sigma_xV_x^T \\
yy^TU_x(\Sigma_x + D\Sigma_x^{-1})^{-1}V_x &= U_x(\Sigma_x + D) \\
yy^TU_x &= U_x(\Sigma_x^2 + D)
\end{align*}$$

which is an eigenvalue problem in the transformed space.

A.2 Derivatives for the Optimization

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

Decompose $K_m = U_mU_m^T$, we have $K = \phi(X)(\otimes_{m=1}^M U_m)(\otimes_{m=1}^M U_m^T)\phi(X)^T$.

Let $\hat{U} = \phi(X)(\otimes_{m=1}^M U_m)$, we have $K = \hat{U}\hat{U}^T$.

The negative log-likelihood

$$L = \frac{1}{2}y^T(\hat{U}\hat{U}^T + D)^{-1}y + \frac{1}{2}\log\det(\hat{U}\hat{U}^T + D) + \text{const}$$

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

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

$$L = \frac{1}{2}D^{-1}y^Ty - \frac{1}{2}D^{-1}y^T\hat{U}\Sigma^{-1}\hat{U}^Ty + \frac{1}{2}\log\det(\Sigma) + \text{const}$$

Take derivative over $U_{m(i,j)}$, we have

$$\frac{\partial L}{\partial U_{m(i,j)}} = \text{tr}[(\frac{\partial L}{\partial \hat{U}})^T(\frac{\partial \hat{U}}{\partial U_{m(i,j)}})], \quad \frac{\partial L}{\partial \hat{U}} = \hat{U}(\Sigma^{-1} + wD^{-1}w^T)^{-1} - yD^{-1}w^T$$

$$\frac{\partial \hat{U}}{\partial U_{m(i,j)}} = \frac{\partial \phi(X)}{\partial U_{m(i,j)}} (U_M \otimes \cdots \otimes U_{m(i,j)} \otimes \cdots \otimes U_1) = \frac{\partial \phi(X)}{\partial U_{m(i,j)}} (U_M \otimes \cdots \otimes O_{m(i,j)} \cdots \otimes U_1)$$

Here $O_{m(i,j)} = e_ie_j^T$ is a matrix with all zeros, but the $(i, j)$th entry as one.

The predictive distribution: $p(y_s| x_s, X, y) \sim N(\mu_s, \sigma_s)$:

$$\mu_s = k(x_s, X)(D^{-1} - D^{-1}\hat{U}(D + \hat{U}\hat{U}^T)^{-1}\hat{U}^T)y$$

$$\sigma_s = k(x_s, x_s) - k(x_s, X)(D^{-1} - D^{-1}\hat{U}(D + \hat{U}\hat{U}^T)^{-1}\hat{U}^T)k(x_s, x_s)$$

Where $\hat{U} = \phi(X)(\otimes_{m=1}^M U_m)$.

A.3 Proof for Proposition 2.1

Consider a 3-mode $T_1 \times T_2 \times T_3$ tensor $W$ of functions $W_{(1)} = [w_1(X), \cdots, w_T(X)]$

$$W = S \times_1 U_1(X) \times_2 U_2 \times_3 U_3$$
where $\mathbf{U}_m$ is an orthogonal $T_m \times R_m$ matrix. Assuming $\mathbf{U}_1(\mathcal{X})$ satisfies $E[\mathbf{U}_1^T \mathbf{U}_1] = \mathbf{I}$ (orthogonal design after rotation).

With Tucker property

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

The population risk can be written as

$$\mathcal{L}(\mathcal{W}) = \text{tr} \left\{ (\mathcal{Y} - \mathcal{X}, \mathcal{W})(\mathcal{Y} - \mathcal{X}, \mathcal{W})^T \right\} = \text{tr} \left\{ \mathbf{S}(1)(\mathbf{U}_2 \mathbf{U}_3)^T \right\} E[\text{cov}(\mathcal{Y}, \mathbf{U}_1(\mathcal{X}))(\mathbf{S}(1)(\mathbf{U}_2 \mathbf{U}_3)^T + E(\mathcal{Y})^T)$$

Denote $E[\text{cov}(\mathcal{Y}, \mathbf{U}_1(\mathcal{X}))] = \mathbf{S}(1)$, bound the difference

$$\mathcal{L}(\mathcal{W}) - \hat{\mathcal{L}}(\mathcal{W}) = \text{tr} \left\{ \mathbf{S}(1)(\mathbf{U}_2 \mathbf{U}_3)^T \right\} \left( \mathbf{S}(1) + \hat{\mathbf{S}}(1) \right) \left( \mathbf{S}(1)(\mathbf{U}_2 \mathbf{U}_3)^T \right)$$

$$\leq \| \mathbf{S}(1)(\mathbf{U}_2 \mathbf{U}_3)^T \|_2 \left\| \mathbf{S}(1) + \hat{\mathbf{S}}(1) \right\|_2$$

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

$$\| \mathbf{A} \mathbf{B} \|_s \leq \| \mathbf{A} \|_p \| \mathbf{B} \|_q \quad 1/p + 1/q = 1$$

Given that $\text{sup}_{\mathbf{U}_1} \| \mathbf{S}(1) - \hat{\mathbf{S}}(1) \|_2 = \mathcal{O}_p \left( \frac{1}{\sqrt{T_3 + \log(T_3 T_2 T_3)}} \right)$

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

$$\mathcal{L}(\hat{\mathcal{W}}) - \mathcal{L}(\mathcal{W}_*) = (\mathcal{L}(\hat{\mathcal{W}}) - \hat{\mathcal{L}}(\hat{\mathcal{W}})) + (\hat{\mathcal{L}}(\mathcal{W}_*) - \mathcal{L}(\mathcal{W}_*))$$

$$\leq |(\mathcal{L}(\hat{\mathcal{W}}) - \hat{\mathcal{L}}(\hat{\mathcal{W}})) - (\mathcal{L}(\mathcal{W}_*) - \hat{\mathcal{L}}(\mathcal{W}_*))$$

$$\leq 2 \text{sup}_{\mathcal{W} \in \mathcal{L}} \{ \mathcal{L}(\mathcal{W}) - \hat{\mathcal{L}}(\mathcal{W}) \}$$

$$\leq \mathcal{O} \left( \| \mathbf{S}(1) \|_2^2 \| \mathbf{S}(1) - \hat{\mathbf{S}}(1) \|_2 \right)$$

if we assume $\| \mathbf{S}(1) \|_2^2 = \mathcal{O} \left( \frac{T_3 + \log(T_3 T_2 T_3)}{N} \right)^{1/4}$, then $\mathcal{L}(\hat{\mathcal{W}}) - \mathcal{L}(\mathcal{W}_*) \leq \mathcal{O}(1)$, 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 Haldes 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:

$$\hat{\epsilon} = \text{tr}(\mathbf{A}^{-1} + \mathbf{P}' \mathbf{D}^{-1} \mathbf{P})^{-1}$$

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

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

$$\mathcal{G}^{-1} = \mathbf{A}^{-1} + \mathbf{P}' \mathbf{D}^{-1} \mathbf{P} + \sum_{t} v_t \mathbf{P}_t$$

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

Evaluating the change

$$\mathcal{G}(n+1) - \mathcal{G}(n) = [\mathcal{G}^{-1}(n) + \sigma_t^{-2} \psi_t \psi_t^T]^{-1} - \mathcal{G}(n) = \frac{\mathcal{G}(n) \psi_t \psi_t^T \mathcal{G}(n)}{\sigma_t^2 + \psi_t^T \mathcal{G}(n) \psi_t}$$
where element \((\psi_t)_i = \delta_{\tau_{n+1},i} \phi_{it}(x_{n+1})\) and \(\tau\) maps the global sample index to task-specific sample index. Introducing \(G = E_D[\mathcal{G}]\) and take expectation over numerator and denominator separately, we have

\[
\frac{\partial G}{\partial n_t} = -\frac{E_D[\mathcal{G}P_t\mathcal{G}]}{\sigma_t^2 + \text{tr}P_tG}
\]

Since generalization error \(\epsilon_t = \text{tr}P_tG\), we have that \(-E_D[\mathcal{G}P_t\mathcal{G}] = \frac{\partial}{\partial n_t} E_D[\mathcal{G}] = \frac{\partial G}{\tau_t}\). Multiplying \(P_s\) on both sides yields the approximation for the expected change:

\[
\frac{\partial P_s G}{\partial n_t} = \frac{\partial \epsilon_s}{\partial n_t} = \frac{1}{\sigma_t^2 + \epsilon_t} \frac{\partial \epsilon_s}{\partial v_t}
\]

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

\[
\epsilon_t(N) = \text{tr}P_t \left( \Lambda^{-1} + \sum_s \frac{n_s}{\sigma_s^2 + \epsilon_s} \right)^{-1}
\]