6 Appendix

6.1 Proof of Lemma 3.2

First notice that under the misspecified model in (15), the likelihood is given by

$$p(\{y_i,\mathbf{t}_i\}_{i=1}^n \,|\, \boldsymbol{\nu}) = \prod_{i=1}^n p((y_i,\mathbf{t}_i)\,|\,\boldsymbol{\nu}) = \prod_{i=1}^n p(y_i\,|\,\mathbf{t}_i,\boldsymbol{\nu})p(\mathbf{t}_i) = \frac{1}{\sigma_\epsilon^n}\prod_{i=1}^n p(\mathbf{t}_i)\phi\left(\frac{y_i-\boldsymbol{\nu}^T\mathbf{t}_i}{\sigma_\epsilon}\right), \tag{29}$$

where $\phi(z)$ denotes the standard normal density evaluated at z, and it is easy to find the (misspecified) posterior distribution

$$\begin{aligned} p_M(\boldsymbol{\nu}\,|\,\{y_i,\mathbf{t}_i\}_{i=1}^n) = \frac{p(\{y_i,\mathbf{t}_i\}_{i=1}^n\,|\,\boldsymbol{\nu})p(\boldsymbol{\nu})}{p(\{y_i,\mathbf{x}_i\}_{i=1}^n)} &\propto \phi\left(\frac{\|\boldsymbol{\nu}\|}{\sigma}\right)\prod_{i=1}^n \phi\left(\frac{y_i-\boldsymbol{\nu}^T\mathbf{t}_i}{\sigma_\epsilon}\right) \\ &= \phi\left(\frac{\|\boldsymbol{\nu}\|}{\sigma}\right)\phi\left(\frac{\left\|\mathbf{y}-\boldsymbol{\nu}^T\mathbf{t}\right\|}{\sigma_\epsilon}\right). \end{aligned} \tag{30}$$

Now, noticing

$$\frac{\|\boldsymbol{\nu}\|^2}{\sigma^2} + \frac{\left\|\mathbf{y}-\boldsymbol{\nu}^T\mathbf{t}\right\|^2}{\sigma_\epsilon^2} = \boldsymbol{\nu}^T\left(\frac{1}{\sigma^2}+\sum_{i=1}^n\frac{\mathbf{t}_i\mathbf{t}_i^T}{\sigma_\epsilon^2}\right)\boldsymbol{\nu} - 2\boldsymbol{\nu}^T\left(\sum_{i=1}^n\frac{y_i\mathbf{t}_i^T}{\sigma_\epsilon^2}\right),$$

then denoting

$$\begin{aligned} \boldsymbol{\Sigma}_M &= \left(\frac{1}{\sigma^2}\mathbb{I}_{d\times d} + \frac{1}{\sigma_\epsilon^2}\sum_{i=1}^n \mathbf{t}_i\mathbf{t}_i^T\right)^{-1} \\ \boldsymbol{\mu}_M &= \boldsymbol{\Sigma}_M\left(\frac{1}{\sigma_\epsilon^2}\sum_{i=1}^n y_i\mathbf{t}_i\right), \end{aligned} \tag{31}$$

we can see from (30) that

$$p_M(\boldsymbol{\nu}\,|\,\{y_i,\mathbf{t}_i\}_{i=1}^n) \sim \mathcal{N}\left(\boldsymbol{\mu}_M,\boldsymbol{\Sigma}_M\right). \tag{32}$$

Similarly, if we were instead to use the true likelihood with the prior given in (20), we would find,

$$p_T((\boldsymbol{\nu},\boldsymbol{\gamma})\,|\,\{y_i,\mathbf{t}_i,\mathbf{x}_i\}_{i=1}^n) \sim \mathcal{N}\left(\boldsymbol{\mu}_T^{full},\boldsymbol{\Sigma}_T^{full}\right), \tag{33}$$

where

$$\begin{aligned} \boldsymbol{\Sigma}_T^{full} &= \left(\frac{1}{\sigma^2}\mathbb{I}_{(d+k)\times(d+k)} + \frac{1}{\sigma_\epsilon^2}\sum_{i=1}^n(\mathbf{t}_i,\mathbf{x}_i)(\mathbf{t}_i,\mathbf{x}_i)^T\right)^{-1} \\ \boldsymbol{\mu}_T^{full} &= \boldsymbol{\Sigma}_T^{full}\left(\frac{1}{\sigma_\epsilon^2}\sum_{i=1}^n y_i(\mathbf{t}_i,\mathbf{x}_i)\right). \end{aligned} \tag{34}$$

Integrating out $\boldsymbol{\gamma}$ from (33) we have

$$p_T(\boldsymbol{\nu} \,|\, \{y_i, \mathbf{t}_i, \mathbf{x}_i\}_{i=1}^n) \sim \mathcal{N}\left(\boldsymbol{\mu}_T, \boldsymbol{\Sigma}_T\right), \tag{35}$$

where, using $\mathbf{J} \in \mathbb{R}^{d\times(d+k)}$ to denote a matrix that selects the first d elements when applied to a vector, i.e. it is a concatenation of the $d \times d$ identity matrix with a $d \times k$ matrix of all zeros,

$$\begin{aligned} \boldsymbol{\Sigma}_T &= \mathbf{J}\left(\frac{1}{\sigma^2}\mathbb{I}_{(d+k)\times(d+k)} + \frac{1}{\sigma_\epsilon^2}\sum_{i=1}^n (\mathbf{t}_i, \mathbf{x}_i)(\mathbf{t}_i, \mathbf{x}_i)^T\right)^{-1}\mathbf{J}^T \\ \boldsymbol{\mu}_T &= \mathbf{J}\boldsymbol{\Sigma}_T^{full}\left(\frac{1}{\sigma_\epsilon^2}\sum_{i=1}^n y_i(\mathbf{t}_i, \mathbf{x}_i)\right). \end{aligned} \tag{36}$$

Next we consider the β-posterior coming from the β-tempered likelihood in the misspecified case, i.e. $p_{M,\beta}(\boldsymbol{\nu} \,|\, \{y_i, \mathbf{t}_i\}_{i=1}^n) \propto (p_M(\{y_i, \mathbf{t}_i\}_{i=1}^n \,|\, \boldsymbol{\nu}))^{1/\beta} p(\boldsymbol{\nu})$. Using steps as in (16)-(32), we find that

$$p_{M,\beta}(\boldsymbol{\nu} \,|\, \{y_i, \mathbf{t}_i\}_{i=1}^n) \sim \mathcal{N}\left(\boldsymbol{\mu}_{M,\beta}, \boldsymbol{\Sigma}_{M,\beta}\right), \tag{37}$$

where,

$$\begin{aligned} \boldsymbol{\Sigma}_{M,\beta} &= \left(\frac{1}{\sigma^2}\mathbb{I}_{d\times d} + \frac{1}{\beta\sigma_\epsilon^2}\sum_{i=1}^n \mathbf{t}_i\mathbf{t}_i^T\right)^{-1} \\ \boldsymbol{\mu}_{M,\beta} &= \boldsymbol{\Sigma}_{M,\beta}\left(\frac{1}{\beta\sigma_\epsilon^2}\sum_{i=1}^n y_i\mathbf{t}_i\right), \end{aligned} \tag{38}$$

Now we are able to characterize the β mean field approximation and the usual ($\beta = 1$) mean field approximation calculated with respect to the misspecified posterior in (32). In particular, using (7) work similar to that in (5), we would like to calculate

$$\begin{aligned} q_{MF}^*(\boldsymbol{\nu};\beta) &= \underset{q(\boldsymbol{\nu})\in\mathcal{Q}_{\mathsf{MF}}}{\arg\max}\left\{\int q(\boldsymbol{\nu})\log p_M(\{y_i, \mathbf{t}_i\}_{i=1}^n|\boldsymbol{\nu})d\boldsymbol{\nu} - \beta D_{\mathsf{KL}}(q(\boldsymbol{\nu})||p(\boldsymbol{\nu}))\right\} \\ &= \beta \underset{q(\boldsymbol{\nu})\in\mathcal{Q}_{\mathsf{MF}}}{\arg\max}\left\{D_{\mathsf{KL}}\left(q(\boldsymbol{\nu})||p_{M,\beta}(\boldsymbol{\nu} \,|\, \{y_i, \mathbf{t}_i\}_{i=1}^n)\right)\right\}. \end{aligned}$$

In the above, $p_{M,\beta}(\boldsymbol{\nu} \,|\, \{y_i, \mathbf{t}_i\}_{i=1}^n)$ is given in (37). Now it is straightforward to verify that the mean field distribution $q(\boldsymbol{\nu}) \in \mathcal{Q}_{\mathsf{MF}}$ minimizing the above is going to be Gaussian with mean the same as in (40) but with a variance matrix that is the diagonal of that in (40). Namely,

$$q_{MF}^*(\boldsymbol{\nu};\beta) \sim \mathcal{N}\left(\boldsymbol{\mu}_{MF,\beta}, \boldsymbol{\Sigma}_{MF,\beta}\right), \tag{39}$$

where $\mathbf{\Sigma}_{MF,\beta}$ is a diagonal matrix with, for $i \in \{1, 2, \ldots, d\}$,

$$\begin{aligned}
[\mathbf{\Sigma}_{MF,\beta}]_{ii} &= \left[\left(\frac{1}{\sigma^2}\mathbb{I}_{d\times d} + \frac{1}{\beta\sigma_\epsilon^2}\sum_{i=1}^{n}\mathbf{t}_i\mathbf{t}_i^T\right)^{-1}\right]_{ii} \\
\boldsymbol{\mu}_{MF,\beta} &= \mathbf{\Sigma}_{M,\beta}\left(\frac{1}{\beta\sigma_\epsilon^2}\sum_{i=1}^{n} y_i\mathbf{t}_i\right),
\end{aligned} \tag{40}$$

Having characterized the relevant posterior distributions, we are ready to study the ratios in (11)-(12). Let us first consider (11). Using Lemma 3.1 along with (35) and (39), we find

$$\begin{aligned}
&\frac{D_{\mathsf{KL}}\left(p_T\left(\mathbf{z}|\mathbf{x}\right)\ ||\ q_{MF}^*\left(\mathbf{z};\beta\right)\right)}{D_{\mathsf{KL}}\left(p_T\left(\mathbf{z}|\mathbf{x}\right)\ ||\ q_{MF}^*\left(\mathbf{z};\beta=1\right)\right)} \\
&= \frac{\log\left(\frac{\mathsf{det}(\mathbf{\Sigma}_{MF,\beta})}{\mathsf{det}(\mathbf{\Sigma}_T)}\right) + \mathsf{tr}\left(\mathbf{\Sigma}_T^{-1}\mathbf{\Sigma}_{MF,\beta}\right) + (\boldsymbol{\mu}_{MF,\beta} - \boldsymbol{\mu}_T)^T\mathbf{\Sigma}_{MF,\beta}^{-1}(\boldsymbol{\mu}_{MF,\beta} - \boldsymbol{\mu}_T) - d}{\log\left(\frac{\mathsf{det}(\mathbf{\Sigma}_{MF,1})}{\mathsf{det}(\mathbf{\Sigma}_T)}\right) + \mathsf{tr}\left(\mathbf{\Sigma}_T^{-1}\mathbf{\Sigma}_{MF,1}\right) + (\boldsymbol{\mu}_{MF,1} - \boldsymbol{\mu}_T)^T\mathbf{\Sigma}_{MF,1}^{-1}(\boldsymbol{\mu}_{MF,1} - \boldsymbol{\mu}_T) - d}
\end{aligned} \tag{41}$$

Considering the above, we will study the asymptotic properties of three terms normalized by $1/n$:

1. $\frac{1}{n}\log\left(\frac{\mathsf{det}(\mathbf{\Sigma}_{MF,\beta})}{\mathsf{det}(\mathbf{\Sigma}_T)}\right) - \frac{d}{n}$
2. $\frac{1}{n}\mathsf{tr}\left(\mathbf{\Sigma}_T^{-1}\mathbf{\Sigma}_{MF,\beta}\right)$
3. $\frac{1}{n}(\boldsymbol{\mu}_{MF,\beta} - \boldsymbol{\mu}_T)^T\mathbf{\Sigma}_{MF,\beta}^{-1}(\boldsymbol{\mu}_{MF,\beta} - \boldsymbol{\mu}_T)$

We will argue that terms (1.) and (2.) both converge in probability to 0, therefore the only relevant term in analyzing the limit of (41) is term (3.).

First, we consider term (1.). We show that $\frac{\mathsf{det}(\mathbf{\Sigma}_{MF,\beta})}{\mathsf{det}(\mathbf{\Sigma}_T)}$ is bounded above as $n \to \infty$ and thus term (1.) converges to 0. By the weak law of large numbers and the continuous mapping theorem, we first notice that for $\mathbf{\Sigma}_{MF,\beta}$ given in (40) and $\mathbf{\Sigma}_T$ given in (36), we have

$$\begin{aligned}
n\mathbf{\Sigma}_{MF,\beta} &= \mathrm{diag}\left(\left(\frac{1}{n\sigma^2}\mathbb{I}_{d\times d} + \frac{1}{n\beta\sigma_\epsilon^2}\sum_{i=1}^{n}\mathbf{t}_i\mathbf{t}_i^T\right)^{-1}\right) \xrightarrow{p} \beta\sigma_\epsilon^2\mathrm{diag}\left(\left(\mathbb{E}\left\{\mathbf{t}\mathbf{t}^T\right\}\right)^{-1}\right), \\
n\mathbf{\Sigma}_T &= \mathbf{J}\left(\frac{1}{n\sigma^2}\mathbb{I}_{(d+k)\times(d+k)} + \frac{1}{n\sigma_\epsilon^2}\sum_{i=1}^{n}(\mathbf{t}_i,\mathbf{x}_i)(\mathbf{t}_i,\mathbf{x}_i)^T\right)^{-1}\mathbf{J}^T \xrightarrow{p} \sigma_\epsilon^2\mathbf{J}\left(\mathbb{E}\left\{(\mathbf{t},\mathbf{x})(\mathbf{t},\mathbf{x})^T\right\}\right)^{-1}\mathbf{J}^T.
\end{aligned} \tag{42}$$

Next we notice that $(\mathbb{E}\{(\mathbf{t},\mathbf{x})(\mathbf{t},\mathbf{x})^T\})^{-1}$ is a block matrix of the form $\begin{bmatrix}\mathbb{E}\{\mathbf{t}\mathbf{t}^T\} & \mathbb{E}\{\mathbf{t}\mathbf{x}^T\} \\ \mathbb{E}\{\mathbf{x}\mathbf{t}^T\} & \mathbb{E}\{\mathbf{x}\mathbf{x}^T\}\end{bmatrix}$. Next we recall that when we invert a block matrix,

$$\begin{bmatrix}\mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D}\end{bmatrix} = \begin{bmatrix}(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} & (\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}\mathbf{B}\mathbf{D}^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} & \mathbf{D}^{-1}+\mathbf{D}^{-1}\mathbf{C}(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}\mathbf{B}\mathbf{D}^{-1}\end{bmatrix},$$