# 1 FITC

## 1.1 FITC

$$
\begin{aligned}
q_{\text{FITC}}(\mathbf{f}_* \mid \mathbf{y}) &\sim \mathcal{N}\left(Q_{*\mathbf{f}}(Q_{\mathbf{ff}}+\Lambda)^{-1}\mathbf{y}, K_{**} - Q_{*\mathbf{f}}(Q_{\mathbf{ff}}+\Lambda)^{-1}Q_{\mathbf{f}*}\right) && (1)\\
&\sim \mathcal{N}\left(K_{*\mathbf{u}}\Sigma K_{\mathbf{uf}}\Lambda^{-1}\mathbf{y}, K_{**} - Q_{**} + K_{*\mathbf{u}}\Sigma K_{\mathbf{u}*}\right) && (2)
\end{aligned}
$$

where $\Lambda = \operatorname{diag}\left[K_{\mathbf{ff}} - Q_{\mathbf{ff}} + \sigma_n^2 I\right]$ and $\Sigma = (K_{\mathbf{uu}} + K_{\mathbf{uf}}\Lambda^{-1}K_{\mathbf{fu}})^{-1}$. Here Equation (1) follows from natural derivation and the Equation (2) is more computationally attractive. To see the equivalence, we first apply Woodbury inversion formula,

$$
(\Lambda + Q_{\mathbf{ff}})^{-1} = \Lambda^{-1} - \Lambda^{-1}K_{\mathbf{fu}}(K_{\mathbf{uu}} + K_{\mathbf{uf}}\Lambda^{-1}K_{\mathbf{fu}})^{-1}K_{\mathbf{uf}}\Lambda^{-1} \tag{3}
$$

By definition of $\Sigma$, we have $(\Lambda + Q_{\mathbf{ff}})^{-1} = \Lambda^{-1} - \Lambda^{-1}K_{\mathbf{fu}}\Sigma K_{\mathbf{uf}}\Lambda^{-1}$. Therefore,

$$
\begin{aligned}
\mu_{\mathbf{f}_*} &= Q_{*\mathbf{f}}(Q_{\mathbf{ff}}+\Lambda)^{-1}\mathbf{y} \\
&= K_{*\mathbf{u}}K_{\mathbf{uu}}^{-1}K_{\mathbf{u}*}(\Lambda^{-1} - \Lambda^{-1}K_{\mathbf{fu}}\Sigma K_{\mathbf{uf}}\Lambda^{-1})\mathbf{y} && \text{(Woodbury with defn } \Sigma) \\
&= K_{*\mathbf{u}}K_{\mathbf{uu}}^{-1}(\Sigma^{-1} - K_{\mathbf{uf}}\Lambda^{-1}K_{\mathbf{fu}})\Sigma K_{\mathbf{uf}}\Lambda^{-1}\mathbf{y} && \text{(pull out } \Sigma K_{\mathbf{uf}}\Lambda^{-1}) \\
&= K_{*\mathbf{u}}K_{\mathbf{uu}}^{-1}K_{\mathbf{uu}}\Sigma K_{\mathbf{uf}}\Lambda^{-1}\mathbf{y} && \text{(defn } \Sigma) \\
&= K_{*\mathbf{u}}\Sigma K_{\mathbf{uf}}\Lambda^{-1}\mathbf{y} \\
\text{Cov}_{\mathbf{f}_*} &= K_{**} - Q_{*\mathbf{f}}(Q_{\mathbf{ff}}+\Lambda)^{-1}Q_{\mathbf{f}*} \\
&= K_{**} - K_{*\mathbf{u}}\Sigma K_{\mathbf{uf}}\Lambda^{-1}K_{\mathbf{fu}}K_{\mathbf{uu}}^{-1}K_{\mathbf{u}*} && \text{(results from } \mu_{\mathbf{f}_*}) \\
&= K_{**} - K_{*\mathbf{u}}\Sigma(\Sigma^{-1} - K_{\mathbf{uu}})K_{\mathbf{uu}}^{-1}K_{\mathbf{u}*} && \text{(defn } \Sigma) \\
&= K_{**} - Q_{**} + K_{*\mathbf{u}}\Sigma K_{\mathbf{u}*}
\end{aligned}
$$

## 1.2 Implementation Details

To optimize for inducing variables, we need a computationally stable way to compute $\log p(\mathbf{y} \mid X)$. More specifically, we need a way to compute $\mathbf{y}^T(\Lambda + Q_{\mathbf{ff}})^{-1}\mathbf{y}$ and $\log|\Lambda + Q_{\mathbf{ff}}|$. Referencing (gpflow, a blog),

$$
\begin{aligned}
(\Lambda + Q_{\mathbf{ff}})^{-1} &= (\Lambda + V^TV)^{-1} && (\texttt{chol } K_{\mathbf{uu}} = L_{\mathbf{uu}}L_{\mathbf{uu}}^T,\ \texttt{backsolve } V = L_{\mathbf{uu}}^{-1}K_{\mathbf{uf}},\ Q_{\mathbf{ff}} = V^TV) \\
&= \Lambda^{-1} - \Lambda^{-1}V^T(I + V\Lambda^{-1}V^T)^{-1}V\Lambda^{-1} && \text{(Woodbury)} \\
&= \Lambda^{-1} - \Lambda^{-1}V^TL_B^{-T}L_B^{-1}V\Lambda^{-1} && (B := I + V\Lambda^{-1}V^T,\ \texttt{chol } B = L_BL_B^T) \\
\mathbf{y}^T(\Lambda + Q_{\mathbf{ff}})^{-1}\mathbf{y} &= \mathbf{y}^T\Lambda^{-1}\mathbf{y} - \boldsymbol{\gamma}^T\boldsymbol{\gamma} && (\texttt{backsolve } \boldsymbol{\gamma} = L_B^{-1}V\Lambda^{-1}\mathbf{y}) \\
\log\det(\Lambda + Q_{\mathbf{ff}}) &= \log\det(I + V\Lambda^{-1}V^T) + \log\det(\Lambda) && (\text{ Matrix inversion lemma, } Q_{\mathbf{ff}} = V^TV) \\
&= \log\det(B) + \log\det(\Lambda) && \text{(defn } B) \\
&= 2\sum_{i=1}^m \log\,[L_B]_{ii} + \sum_{i=1}^n \log\,[\Lambda]_{ii}
\end{aligned}
$$

Put everything together, we have an expression that involves mostly $\mathcal{O}(m^3)$ `chol` and $\mathcal{O}(nm^2)$ `backsolve`,

$$
\log p(\mathbf{y} \mid X) = -\frac{1}{2}\mathbf{y}^T\Lambda^{-1}\mathbf{y} + \frac{1}{2}\boldsymbol{\gamma}^T\boldsymbol{\gamma} - \sum_{i=1}^m [L_B]_{ii} - \frac{1}{2}\sum_{i=1}^n [\Lambda]_{ii} - \frac{n}{2}\log(2\pi) \tag{4}
$$

We can also compute the predictive distribution in Equation (2) as follows,

$$
\begin{aligned}
\Sigma &= (K_{\mathbf{uu}} + K_{\mathbf{uf}}\Lambda^{-1}K_{\mathbf{fu}})^{-1} \\
&= (L_{\mathbf{uu}}L_{\mathbf{uu}}^{T} + L_{\mathbf{uu}}L_{\mathbf{uu}}^{-1}K_{\mathbf{uf}}\Lambda^{-1}K_{\mathbf{fu}}L_{\mathbf{uu}}^{-T}L_{\mathbf{uu}}^{T}) && (K_{\mathbf{uu}} = L_{\mathbf{uu}}L_{\mathbf{uu}}^{T}) \\
&= (L_{\mathbf{uu}}BL_{\mathbf{uu}}^{T})^{-1} && (B = I + L_{\mathbf{uu}}^{T}L_{\mathbf{uf}}\Lambda^{-1}K_{\mathbf{fu}}L_{\mathbf{uu}}^{-T}) \\
&= L_{\mathbf{uu}}^{-T}L_{B}^{-T}L_{B}^{-1}L^{-1} \\
\mu_{\mathbf{f}_*} &= K_{*\mathbf{u}}\Sigma K_{\mathbf{uf}}\Lambda^{-1}\mathbf{y} \\
&= K_{*\mathbf{u}}L_{\mathbf{uu}}^{-T}L_{B}^{-T}L_{B}^{-1}L^{-1}K_{\mathbf{uf}}\Lambda^{-1}\mathbf{y} && (\text{ substitute } \Sigma) \\
&= K_{*\mathbf{u}}L_{\mathbf{uu}}^{-T}L_{B}^{-T}\boldsymbol{\gamma} && (\ \boldsymbol{\gamma} = L_{B}^{-1}L_{\mathbf{uu}}^{-1}K_{\mathbf{uf}}\Lambda^{-1}\mathbf{y}) \\
&= \boldsymbol{\omega}^{T}L_{B}^{-T}\boldsymbol{\gamma} && (\texttt{backsolve } \boldsymbol{\omega} = L_{\mathbf{uu}}^{-1}K_{\mathbf{u}*} \text{ implies } Q_{**} = \boldsymbol{\omega}^{T}\boldsymbol{\omega}) \\
\mathrm{Cov}_{\mathbf{f}_*} &= K_{**} - Q_{**} + K_{*\mathbf{u}}\Sigma K_{\mathbf{u}*} \\
&= K_{**} - Q_{**} + K_{*\mathbf{u}}L_{\mathbf{uu}}^{-1}L_{B}^{-1}L_{B}^{-1}L_{\mathbf{uu}}^{-1}K_{\mathbf{u}*} && (\text{substitute } \Sigma) \\
&= K_{**} - \boldsymbol{\omega}^{T}\boldsymbol{\omega} + \boldsymbol{\nu}^{T}\boldsymbol{\nu} && (\texttt{backsolve } \boldsymbol{\nu} = L_{B}^{-1}\boldsymbol{\omega})
\end{aligned}
$$

Note that there are two additional `backsolve` to compute the predictive distribution.