Supplemental Material for Understanding Probabilistic Sparse Gaussian Process Approximations

Matthias Bauer[†‡] **Mark van der Wilk**[†] **Carl Edward Rasmussen**[†]
[†]Department of Engineering, University of Cambridge, Cambridge, UK
[‡]Max Planck Institute for Intelligent Systems, Tübingen, Germany
{msb55, mv310, cer54}@cam.ac.uk

1 Proofs for additional inducing inputs

In this section we give proofs to the claims on how the objective functions for FITC and VFE change upon adding an inducing input. For this, we first restate the objective function:

$$\mathcal{F} = \frac{N}{2}\log(2\pi) + \underbrace{\frac{1}{2}\log|Q_{\mathbf{ff}} + G|}_{\text{complexity penalty}} + \underbrace{\frac{1}{2}\mathbf{y}^\intercal(Q_{\mathbf{ff}} + G)^{-1}\mathbf{y}}_{\text{data fit}} + \underbrace{\frac{1}{2\sigma_n^2}\operatorname{tr}(T)}_{\text{trace term}}, \tag{1}$$

where

$$G_{\text{FITC}} = \operatorname{diag}[K_{\mathbf{ff}} - Q_{\mathbf{ff}}] + \sigma_n^2 I \qquad G_{\text{VFE}} = \sigma_n^2 I \tag{2}$$

$$T_{\text{FITC}} = 0 \qquad T_{\text{VFE}} = K_{\mathbf{ff}} - Q_{\mathbf{ff}}. \tag{3}$$

$K_{\mathbf{ff}}$ denotes the covariance matrix and $Q_{\mathbf{ff}} = K_{\mathbf{fu}}K_{\mathbf{uu}}^{-1}K_{\mathbf{uf}}$ the approximate covariance matrix. Note that the approximate covariance matrix $Q_{\mathbf{ff}}$ is the only quantity depending on the inducing inputs.

The main results we show in this supplement are:

1. Adding an inducing input corresponds to a rank 1 update of the approximate covariance matrix $Q_{\mathbf{ff}} = K_{\mathbf{fu}}K_{\mathbf{uu}}^{-1}K_{\mathbf{uf}}$ (see Section 1.1)
2. For VFE the objective function can never get worse by adding an inducing input anywhere in the input domain (see Section 1.2)
3. For FITC the heteroscedastic noise always decreases when adding an inducing input anywhere in the input domain (see Section 1.3)
4. Adding an inducing input on top of an existing inducing input does not change the objective function, neither for FITC nor for VFE (see Section 1.4)

1.1 Adding an inducing input anywhere

In this section we show that adding an inducing input corresponds to a rank 1 update of the $Q_{\mathbf{ff}}$ matrix for the case without jitter.
Let $K_{\mathbf{uu}}$ denote the covariance matrix of the M inducing inputs. We then add a new $M+1$st inducing input and denote all quantities depending on the new set of $M+1$ inducing inputs by a superscript $+$.
The updated approximate covariance matrix $Q_{\mathbf{ff}}^+$ is then given by:

$$Q_{\mathbf{ff}}^+ = K_{\mathbf{fu}}^+(K_{\mathbf{uu}}^+)^{-1}K_{\mathbf{uf}}^+ \tag{4}$$

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

We proceed by first computing an explicit expression for the $M+1 \times M+1$ matrix $(K^+_{\mathbf{uu}})^{-1}$ before computing $Q^+_{\mathbf{ff}}$. For this we employ the block matrix inversion formula[1], see Eq. (8)

$$
\begin{aligned}
(K^+_{\mathbf{uu}})^{-1} &= \begin{pmatrix} K_{\mathbf{uu}} & k_{\mathbf{u}} \\ k_{\mathbf{u}}^\intercal & k \end{pmatrix}^{-1} \\
&= \begin{pmatrix} K_{\mathbf{uu}}^{-1} + \frac{1}{c}\mathbf{a}\mathbf{a}^\intercal & -\frac{1}{c}\mathbf{a} \\ -\frac{1}{c}\mathbf{a}^\intercal & \frac{1}{c} \end{pmatrix} \qquad\qquad \mathbf{a} = K_{\mathbf{uu}}^{-1} k_{\mathbf{u}}
\end{aligned}
$$

where $c = k - k_{\mathbf{u}}^\intercal K_{\mathbf{uu}}^{-1} k_{\mathbf{u}}$ is the Schur complement of $K_{\mathbf{uu}}$, $k_{\mathbf{u}}^\intercal = (k(Z_1, Z_{M+1}), \ldots, k(Z_M, Z_{M+1}))$ is the vector of covariances between the old inducing inputs and the added inducing input and $k = k(Z_{M+1}, Z_{M+1})$ is the covariance function evaluated at the new inducing input. Note that c needs to be non-zero for this expression to make sense.

$$
K^+_{\mathbf{uf}} = \begin{pmatrix} K_{\mathbf{uf}} \\ k_{\mathbf{f}}^\intercal \end{pmatrix}
$$

where $k_{\mathbf{f}}^\intercal = (k(Z_{M+1}, x_1), \ldots, k(Z_{M+1}, x_N))$ is the vector of covariances between the data points and the new inducing input.
We can now compute $Q^+_{\mathbf{ff}}$ by the product in Eq. (4) to see that is is indeed given by a rank 1 update of $Q_{\mathbf{ff}}$

$$
\begin{aligned}
Q^+_{\mathbf{ff}} &= K^+_{\mathbf{fu}} (K^+_{\mathbf{uu}})^{-1} K^+_{\mathbf{uf}} \\
&= K_{\mathbf{fu}} K_{\mathbf{uu}}^{-1} K_{\mathbf{uf}} + \frac{1}{c} \big(K_{\mathbf{fu}} \mathbf{a}\mathbf{a}^\intercal K_{\mathbf{uf}} + K_{\mathbf{fu}} \mathbf{a}\mathbf{a}^\intercal K_{\mathbf{uf}} \\
&\qquad - K_{\mathbf{fu}} \mathbf{a} k_{\mathbf{f}}^\intercal - k_{\mathbf{f}} \mathbf{a}^\intercal K_{\mathbf{uf}} + k_{\mathbf{f}} k_{\mathbf{f}}^\intercal \big) \\
&= K_{\mathbf{fu}} K_{\mathbf{uu}}^{-1} K_{\mathbf{uf}} + \frac{1}{c} \left(K_{\mathbf{fu}} \mathbf{a} - k_{\mathbf{f}} \right) \left(K_{\mathbf{fu}} \mathbf{a} - k_{\mathbf{f}} \right)^\intercal \\
&= K_{\mathbf{fu}} K_{\mathbf{uu}}^{-1} K_{\mathbf{uf}} + \mathbf{b}\mathbf{b}^\intercal \\
&= Q_{\mathbf{ff}} + \mathbf{b}\mathbf{b}^\intercal \\
Q^+_{\mathbf{ff}} &= Q_{\mathbf{ff}} + \mathbf{b}\mathbf{b}^\intercal
\end{aligned} \tag{5}
$$

where we have introduced the rank 1 update vector $\mathbf{b} = \frac{1}{\sqrt{c}}(K_{\mathbf{fu}} K_{\mathbf{uu}}^{-1} k_{\mathbf{u}} - k_{\mathbf{f}})$.

Thus, in the case of no *jitter*, the update is indeed given by a rank 1 update. These results also extend to the case with finite jitter, which is then absorbed into the definition of $K_{\mathbf{uu}}$ and k, respectively.

1.2 The VFE objective function always improves when adding an additional inducing input

We now compute the change in objective function when adding the $M+1$ st inducing input:

$$
\begin{aligned}
2(\mathcal{F}^+ - \mathcal{F}) &= \log \left| Q^+_{\mathbf{ff}} + \sigma_n^2 I \right| - \log \left| Q_{\mathbf{ff}} + \sigma_n^2 I \right| + \mathbf{y}^\intercal (Q^+_{\mathbf{ff}} + \sigma_n^2 I)^{-1} \mathbf{y} - \mathbf{y}^\intercal (Q_{\mathbf{ff}} + \sigma_n^2 I)^{-1} \mathbf{y} \\
&\quad + \frac{1}{\sigma_n^2} \operatorname{tr}(K_{\mathbf{ff}} - Q^+_{\mathbf{ff}}) - \frac{1}{\sigma_n^2} \operatorname{tr}(K_{\mathbf{ff}} - Q_{\mathbf{ff}}) \\
&= \log \left| Q_{\mathbf{ff}} + \mathbf{b}\mathbf{b}^\intercal + \sigma_n^2 I \right| - \log \left| Q_{\mathbf{ff}} + \sigma_n^2 I \right| \\
&\quad + \mathbf{y}^\intercal (Q_{\mathbf{ff}} + \mathbf{b}\mathbf{b}^\intercal + \sigma_n^2 I)^{-1} \mathbf{y} - \mathbf{y}^\intercal (Q_{\mathbf{ff}} + \sigma_n^2 I)^{-1} \mathbf{y} - \frac{1}{\sigma_n^2} \operatorname{tr}(\mathbf{b}\mathbf{b}^\intercal)
\end{aligned}
$$

[1] `https://en.wikipedia.org/wiki/Block_matrix#Block_matrix_inversion`

To deal with the log-determinant-terms and the inverses, we employ the Matrix determinant lemma[2] and the Sherman–Morrison formula[3], respectively

$$
\begin{aligned}
2(\mathcal{F}^+ - \mathcal{F}) &= \log(1 + \mathbf{b}^\intercal (Q_{\mathbf{ff}} + \sigma_n^2 I)^{-1}\mathbf{b}) + \log\left|Q_{\mathbf{ff}} + \sigma_n^2 I\right| - \log\left|Q_{\mathbf{ff}} + \sigma_n^2 I\right| \\
&\quad + \mathbf{y}^\intercal (Q_{\mathbf{ff}} + \sigma_n^2 I)^{-1}\mathbf{y} - \mathbf{y}^\intercal \frac{(Q_{\mathbf{ff}} + \sigma_n^2 I)^{-1}\mathbf{b}\mathbf{b}^T(Q_{\mathbf{ff}} + \sigma_n^2 I)^{-1}}{1 + \mathbf{b}^T(Q_{\mathbf{ff}} + \sigma_n^2 I)^{-1}\mathbf{b}}\mathbf{y} \\
&\quad - \mathbf{y}^\intercal (Q_{\mathbf{ff}} + \sigma_n^2 I)^{-1}\mathbf{y} + \frac{1}{\sigma_n^2}\operatorname{tr}(\mathbf{b}\mathbf{b}^\intercal) \\
&= \log(1 + \mathbf{b}^\intercal (Q_{\mathbf{ff}} + \sigma_n^2 I)^{-1}\mathbf{b}) - \frac{1}{\sigma_n^2}\operatorname{tr}(\mathbf{b}\mathbf{b}^\intercal) \\
&\quad - \mathbf{y}^\intercal \frac{(Q_{\mathbf{ff}} + \sigma_n^2 I)^{-1}\mathbf{b}\mathbf{b}^T(Q_{\mathbf{ff}} + \sigma_n^2 I)^{-1}}{1 + \mathbf{b}^T(Q_{\mathbf{ff}} + \sigma_n^2 I)^{-1}\mathbf{b}}\mathbf{y}
\end{aligned}
$$

We can bound the first two terms by noting

$$
\begin{aligned}
\operatorname{tr}(\mathbf{b}\mathbf{b}^\intercal) &= \mathbf{b}^\intercal\mathbf{b} \\
\log(1 + x) &\leq x \\
\mathbf{b}^\intercal (Q_{\mathbf{ff}} + \sigma_n^2 I)^{-1}\mathbf{b} &\leq \frac{1}{\sigma_n^2}\mathbf{b}^\intercal\mathbf{b}
\end{aligned}
$$

Thus,

$$
\log(1 + \mathbf{b}^\intercal (Q_{\mathbf{ff}} + \sigma_n^2 I)^{-1}\mathbf{b}) - \frac{1}{\sigma_n^2}\operatorname{tr}(\mathbf{b}\mathbf{b}^\intercal) \leq 0
$$

and equality holds for $\mathbf{b} = 0$, as is the case when both inducing inputs lie on top of each other. It remains to show that the term including the $\mathbf{y}$s (including its sign) is non-positive. This can be shown quite easily:

$$
\begin{aligned}
-\mathbf{y}^\intercal \frac{(Q_{\mathbf{ff}} + \sigma_n^2 I)^{-1}\mathbf{b}\mathbf{b}^T(Q_{\mathbf{ff}} + \sigma_n^2 I)^{-1}}{1 + \mathbf{b}^T(Q_{\mathbf{ff}} + \sigma_n^2 I)^{-1}\mathbf{b}}\mathbf{y} &= -\frac{(\mathbf{y}^\intercal (Q_{\mathbf{ff}} + \sigma_n^2 I)^{-1}\mathbf{b})^2}{1 + \mathbf{b}^T(Q_{\mathbf{ff}} + \sigma_n^2 I)^{-1}\mathbf{b}} \\
&\leq -(\mathbf{y}^\intercal (Q_{\mathbf{ff}} + \sigma_n^2 I)^{-1}\mathbf{b})^2 \\
&\leq 0
\end{aligned}
$$

where the second to last inequality holds as $(Q_{\mathbf{ff}} + \sigma_n^2 I)^{-1}$ is positive definite. Equalities hold, again, if $\mathbf{b} = 0$ which corresponds to duplication of an existing inducing input.

This concludes the proof that the VFE objective function always improves or stays the same.

The change of the objective function for FITC is less clear than for VFE. We can give no proofs about the changes in general. In experiments, we observe that the change in the data fit term can be positive or negative, whereas the complexity penalty term seems to always improve by adding an inducing input. We hypothesise that this is indeed the case but cannot give a proof for this claim.

In this section we have assumed that all matrix inverses exist and that for duplication of an inducing input we find $\mathbf{b} = 0$. However, for a duplicate inducing input, the matrix $K_{\mathbf{uu}}$ becomes singular, such that care has to be taken when reasoning. In Section 1.4 we show that these arguments can, indeed, be made rigorous and that $\mathbf{b} = 0$ for duplicate inducing inputs. Thus, when duplicating an inducing input, the VFE and FITC objective functions do not change. For VFE, this configuration corresponds to a maximum of the objective function.

1.3 The heteroscedastic noise is decresed when new inducing inputs are added

While the objective function can change either way, the heteroscedastic noise, which is given by $\operatorname{diag}(K_{\mathbf{ff}} - Q_{\mathbf{ff}})$ always decreases or remains the same when a new inducing input is added:

$$
\begin{aligned}
\operatorname{diag}(K_{\mathbf{ff}} - Q_{\mathbf{ff}}^+) &= \operatorname{diag}(K_{\mathbf{ff}} - (Q_{\mathbf{ff}} + \mathbf{b}\mathbf{b}^\intercal)) && (6) \\
&= \operatorname{diag}(K_{\mathbf{ff}} - Q_{\mathbf{ff}}) - \operatorname{diag}(\mathbf{b}\mathbf{b}^\intercal) && (7)
\end{aligned}
$$

The diagonal elements of $\mathbf{b}\mathbf{b}^\intercal$ are given by b_m^2, which are always larger or equal to zero, such that the heteroscedastic noise always decreases (or stays the same).

[2] https://en.wikipedia.org/wiki/Matrix_determinant_lemma

[3] https://en.wikipedia.org/wiki/Sherman-Morrison_formula

1.4 Adding an inducing input on top of another inducing input

In this section we show that duplication of an inducing input, that is, placing an additional inducing input on top of an existing one, does not change the approximate covariance matrix: $Q^+_{\mathbf{ff}} = Q_{\mathbf{ff}}$.
One might be tempted to use Eq. (5) to evaluate the update when placing an additional inducing input on top of an existing one. In that case $K^{-1}_{\mathbf{uu}} k_{\mathbf{u}} = \widehat{e}_M$, where $\widehat{e}_M$ denotes the indicator vector with a one at the Mth position and zeros otherwise, and $K_{\mathbf{fu}} K^{-1}_{\mathbf{uu}} k_{\mathbf{u}} = k_{\mathbf{f}}$ suggesting $\mathbf{b} = 0$. However, in this case, the Schur complement vanishes as well, $c = 0$.
In the following we show that this reasoning can be made exact by considering a finite *jitter* term ϵI added onto $K^+_{\mathbf{uu}}$ before inversion. The result can be expanded to second order in ϵ and the limit $\epsilon \to 0$ then leads to the desired result. The intuition behind the fact that Eq. (4) is well behaved, even if $K^+_{\mathbf{uu}}$ is singular, is, that the eigenvector of $K^+_{\mathbf{uu}}$ that corresponds to the zero eigenvalue is never excited by the matrix $K^+_{\mathbf{uf}}$ which has a duplicate row. The eigenvector only has two non-zero elements, which have the same absolute value but different signs, thus cancelling with the duplicate rows in $K^+_{\mathbf{uf}}$.
Moreover, we obtain a correction term that scales with the jitter. For reasons of numerical stability, one has to employ some form of regularisation of the (possibly) singular matrix $K_{\mathbf{uu}}$ in practise. One common way that is implemented in many toolboxes is the constant jitter ϵI introduced above. We assume that the original $K_{\mathbf{uu}}$ is non-singular for now.

Similarly to before, we again employ the block matrix inversion formula[4] of the following form:

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} A^{-1} + A^{-1}BF^{-1}CA^{-1} & -A^{-1}BF^{-1} \\ -F^{-1}CA^{-1} & F^{-1} \end{pmatrix} \tag{8}$$

where $F = D - CA^{-1}B$ is the Schur complement of A.

$$(K^+_{\mathbf{uu}} + \epsilon I_{M+1\times M+1})^{-1} = \begin{pmatrix} K_{\mathbf{uu}} + \epsilon I_{M\times M} & k_{\mathbf{u}} \\ k_{\mathbf{u}}^{\intercal} & k + \epsilon \end{pmatrix}^{-1} \tag{9}$$

where $k_{\mathbf{u}}^{\intercal} = (k(Z_1, Z_M), k(Z_2, Z_M), \ldots, k(Z_M, Z_M))$ and $k = k(Z_M, Z_M)$ similarly to before. In order to perform the inversion, we expand the inverses in Eq. (8) to second order in ϵ:

$$\begin{aligned}
(K_{\mathbf{uu}} + \epsilon I)^{-1} &\approx K^{-1}_{\mathbf{uu}}(1 - \epsilon K^{-1}_{\mathbf{uu}} + \epsilon^2 K^{-2}_{\mathbf{uu}} - \epsilon^3 K^{-3}_{\mathbf{uu}}) && A^{-1} \\
(K_{\mathbf{uu}} + \epsilon I)^{-1} k_{\mathbf{u}} &\approx \widehat{e}_M - \epsilon k^{-1}_{\mathbf{u}} + \epsilon^2 K^{-1}_{\mathbf{uu}} k^{-1}_{\mathbf{u}} - \epsilon^3 K^{-2}_{\mathbf{uu}} k^{-1}_{\mathbf{u}} && A^{-1}B \\
k_{\mathbf{u}}^T (K_{\mathbf{uu}} + \epsilon I)^{-1} &\approx \widehat{e}_M^T - \epsilon (k^{-1}_{\mathbf{u}})^T + \epsilon^2 (k^{-1}_{\mathbf{u}})^T K^{-1}_{\mathbf{uu}} - \epsilon^3 (k^{-1}_{\mathbf{u}})^T K^{-2}_{\mathbf{uu}} && CA^{-1} \\
k_{\mathbf{u}}^T (K_{\mathbf{uu}} + \epsilon)^{-1} k_{\mathbf{u}} &\approx k - \epsilon + \epsilon^2 k^{-1} - \epsilon^3 (k^{-1}_{\mathbf{u}})^T k^{-1}_{\mathbf{u}} && CA^{-1}B \\
k + \epsilon - k_{\mathbf{u}}^T (K_{\mathbf{uu}} + \epsilon)^{-1} k_{\mathbf{u}} &\approx 2\epsilon(1 - \frac{\epsilon}{2} k^{-1} + \frac{\epsilon^2}{2} (k^{-1}_{\mathbf{u}})^T k^{-1}_{\mathbf{u}}) && \text{Schur} \\
(k + \epsilon - k_{\mathbf{u}}^T (K_{\mathbf{uu}} + \epsilon)^{-1} k_{\mathbf{u}})^{-1} &\approx \frac{1}{2\epsilon} + \frac{1}{4} k^{-1} - \frac{\epsilon}{4} (k^{-1}_{\mathbf{u}})^T k^{-1}_{\mathbf{u}} + \frac{\epsilon}{8} (k^{-1})^2 && \text{Schur}^{-1}
\end{aligned}$$

where $\widehat{e}_M = (0, \ldots, 0, 1)^{\intercal}$ is the indicator vector with a one at position M and $k^{-1}_{\mathbf{u}} = (k^{-1}(Z_1, Z_M), \ldots, k^{-1}(Z_M, Z_M))^{\intercal}$ is the Mth column of $K^{-1}_{\mathbf{uu}}$. Note that the elements of $k^{-1}_{\mathbf{u}}$ are not element wise inverses but elements of an inverse matrix! Analogously, k^{-1} denotes the (M, M) element of the matrix $K^{-1}_{\mathbf{uu}}$.

[4] https://en.wikipedia.org/wiki/Block_matrix#Block_matrix_inversion

$$
\begin{aligned}
(K^+_{\mathbf{uu}}+\epsilon)^{-1} = &\left(\begin{array}{ccc|c} & K^{-1}_{\mathbf{uu}} & & \begin{matrix}0\\ \vdots\\ 0\end{matrix} \\ \hline 0 & \cdots & 0 & 0 \end{array}\right) - \epsilon \left(\begin{array}{ccc|c} & K^{-2}_{\mathbf{uu}} & & \begin{matrix}0\\ \vdots\\ 0\end{matrix} \\ \hline 0 & \cdots & 0 & 0 \end{array}\right) + \frac{\epsilon}{2}\left(\begin{array}{ccc|c} & k^{-1}_{\mathbf{u}}(k^{-1}_{\mathbf{u}})^{\intercal} & & \begin{matrix}0\\ \vdots\\ 0\end{matrix} \\ \hline 0 & \cdots & 0 & 0 \end{array}\right) \\
&+\frac{1}{2}\left(\begin{array}{ccc|c} & & 0 & 0 \\ & 0_{M-1\times M-1} & \vdots & \vdots \\ & & 0 & 0 \\ 0 & \cdots & 0 \quad \epsilon^{-1} & -\epsilon^{-1} \\ \hline 0 & \cdots & 0 \quad -\epsilon^{-1} & \epsilon^{-1} \end{array}\right) \\
&+\frac{1}{4}\left(\begin{array}{ccc|c} & & 0 & 0 \\ & 0_{M-1\times M-1} & \vdots & \vdots \\ & & 0 & 0 \\ 0 & \cdots & 0 \quad k^{-1} & -k^{-1} \\ \hline 0 & \cdots & 0 \quad -k^{-1} & k^{-1} \end{array}\right) + \frac{1}{2}\left(\begin{array}{cc|c} 0_{M-1\times M-1} & -k^{-1}_{u\backslash m} & k^{-1}_{\mathbf{u}} \\ -(k^{-1}_{u\backslash m})^T & -2k^{-1} & \\ \hline (k^{-1}_{\mathbf{u}})^T & & 0 \end{array}\right) \\
&+\frac{\epsilon}{4}\left(\begin{array}{cc|c} 0_{M-1\times M-1} & 2K^{-1}_{\mathbf{uu}}k^{-1}_{\mathbf{u}} - k^{-1}k^{-1}_{\mathbf{u}} & -(2K^{-1}_{\mathbf{uu}}k^{-1}_{\mathbf{u}} - k^{-1}k^{-1}_{\mathbf{u}}) \\ (2K^{-1}_{\mathbf{uu}}k^{-1}_{\mathbf{u}} - k^{-1}k^{-1}_{\mathbf{u}})^T & & \\ \hline -(2K^{-1}_{\mathbf{uu}}k^{-1}_{\mathbf{u}} - k^{-1}k^{-1}_{\mathbf{u}})^T & & 0 \end{array}\right) \\
&+\frac{\epsilon}{8}\left(\begin{array}{ccc|c} & & 0 & 0 \\ & 0_{M-1\times M-1} & \vdots & \vdots \\ & & 0 & 0 \\ 0 & \cdots & 0 \quad (k^{-1})^2 - 2(k^{-1}_{\mathbf{u}})^T k^{-1}_{\mathbf{u}} & -((k^{-1})^2 - 2(k^{-1}_{\mathbf{u}})^T k^{-1}_{\mathbf{u}}) \\ \hline 0 & \cdots & 0 \quad -((k^{-1})^2 - 2(k^{-1}_{\mathbf{u}})^T k^{-1}_{\mathbf{u}}) & (k^{-1})^2 - 2(k^{-1}_{\mathbf{u}})^T k^{-1}_{\mathbf{u}} \end{array}\right) + \mathcal{O}(\epsilon^2)
\end{aligned}
$$

For the matrix $(K_{\mathbf{uu}}+\epsilon I)^{-1}$ we find:

$$(K_{\mathbf{uu}}+\epsilon)^{-1} = K^{-1}_{\mathbf{uu}} - \epsilon K^{-2}_{\mathbf{uu}} + \mathcal{O}(\epsilon^2)$$

When we now multiply out the product $K^+_{\mathbf{fu}}(K^+_{\mathbf{uu}}+\epsilon I)^{-1}K^+_{\mathbf{uf}}$, we note that $K^+_{\mathbf{uf}}$ will have a duplicate row and $K^+_{\mathbf{fu}}$ will have a duplicate column. Due to this, all terms that have the submatrix $0_{M-1\times M-1}$ in their upper left hand corner cancel. This includes the term that contains the exploding (in the limit $\epsilon \to 0$) inverse jitter ϵ^{-1}, and we are left with:

$$
\begin{aligned}
Q^+_{\mathbf{ff}} = K^+_{\mathbf{fu}}(K^+_{\mathbf{uu}}+\epsilon I)^{-1}K^+_{\mathbf{uf}} &= K_{\mathbf{fu}}K^{-1}_{\mathbf{uu}}K_{\mathbf{uf}} - \epsilon K_{\mathbf{fu}}K^{-2}_{\mathbf{uu}}K_{\mathbf{uf}} + \frac{\epsilon}{2}K_{\mathbf{fu}}k^{-1}_{\mathbf{u}}(k^{-1}_{\mathbf{u}})^{\intercal}K_{\mathbf{uf}} + \mathcal{O}(\epsilon^2) \\
&= Q_{\mathbf{ff}} - \epsilon K_{\mathbf{fu}}K^{-2}_{\mathbf{uu}}K_{\mathbf{uf}} + \frac{\epsilon}{2}K_{\mathbf{fu}}k^{-1}_{\mathbf{u}}(k^{-1}_{\mathbf{u}})^{\intercal}K_{\mathbf{uf}} + \mathcal{O}(\epsilon^2)
\end{aligned}
$$

Such that the correction to the original approximate covariance matrix is given by:

$$Q^+_{\mathbf{ff}} - Q_{\mathbf{ff}} = \frac{\epsilon}{2}K_{\mathbf{fu}}k^{-1}_{\mathbf{u}}(k^{-1}_{\mathbf{u}})^{\intercal}K_{\mathbf{uf}} + \mathcal{O}(\epsilon^2) \tag{10}$$

We can now take the limit $\epsilon \to 0$ as all the "infinities" have cancelled above. For finite jitter, the correction term is again given by a rank-1 update to first order in ϵ.