# Laplace approximation for DeepGP

September 27, 2020

## 1 Laplace Inference with GTG

In this section we get main formulas for implementtion of Laplace approximation. Let suppose noise in covariance matrix:

$$K = K + \sigma^2 \cdot I = G^T G + \sigma^2 \cdot I.$$

Then inverse matrix will be:

$$K^{-1} = (G^T G + \sigma^2 \cdot I)^{-1} = \frac{1}{\sigma^2} I - \frac{1}{\sigma^4} G^T (I_m + \frac{1}{\sigma^2} G G^T)^{-1} G,$$

$$(K^{-1} + W)^{-1} = ((\frac{1}{\sigma^2} I + W) - \frac{1}{\sigma^4} G^T (I_m + \frac{1}{\sigma^2} G G^T)^{-1} G)^{-1}.$$

Using matrix inversion lemma one more time and assuming

$$W^*_{inv} = (\frac{1}{\sigma^2} I + W)^{-1},$$

we get final formula:

$$(K^{-1} + W)^{-1} = W^*_{inv} + \frac{1}{\sigma^4} W^*_{inv} G^T (I_m + \frac{1}{\sigma^2} G G^T - \frac{1}{\sigma^4} G W^*_{inv} G^T)^{-1} G W^*_{inv}.$$

Posterior - is Gauissian distribution with covariance $(K^{-1} + W)^{-1}$ and mean, which comes from iterations:

$$f_{new} = (K^{-1} + W)^{-1} (W f + \nabla \log(p(y|f)).$$

## 2 Factoring out sigma

Throught the experiments we got some huge value in $W$, which is second derivarive of f. One of obvious ways to battle this problem is to factor out constants from matricies. Again we have covariance matrix:

$$K = K + \sigma^2 \cdot I = QR + \sigma^2 \cdot I.$$

Then inverse matrix will be:

$$K^{-1} = (QR + \sigma^2 \cdot I)^{-1} = \frac{1}{\sigma^2} I - \frac{1}{\sigma^4} Q (I_m + \frac{1}{\sigma^2} RQ)^{-1} R.$$

Posterior therefore looks like that:

$$(\sigma^2 (K^{-1} + W))^{-1} = (I + \sigma^2 W - \frac{1}{\sigma^2} Q (I_m + \frac{1}{\sigma^2} RQ)^{-1} R))^{-1}.$$

Remember that multiplication factor changes its position in case of inversion

$$\frac{1}{\sigma^2}(K^{-1}+W)^{-1} = (\sigma^2(K^{-1}+W))^{-1}.$$

Using matrix inversion lemma one more time and assuming

$$W^*_{inv} = (I+\sigma^2 W)^{-1}$$

we get final formula:

$$\frac{1}{\sigma^2}(K^{-1}+W)^{-1} = W^*_{inv} + \frac{1}{\sigma^2}W^*_{inv}Q(I_m + \frac{1}{\sigma^2}RQ - \frac{1}{\sigma^2}RW^*_{inv}Q)^{-1}RW^*_{inv}.$$

$$(K^{-1}+W)^{-1} = \sigma^2 W^*_{inv} + W^*_{inv}Q(I_m + \frac{1}{\sigma^2}RQ - \frac{1}{\sigma^2}RW^*_{inv}Q)^{-1}RW^*_{inv}.$$

This approach of multiplication by $\sigma^2$ seems to decrease values inside W matricies. However, it only makes things worse on practice through inevitably creating singular matrix during inversion of the middle factor.

# 3 Marginal log likelihood

According to Rasmussen marginal log likelihood is:

$$\log(y|X,\theta) = -\frac{1}{2}\hat{f}^T K^{-1}\hat{f} + \log p(y|\hat{f}) - \frac{1}{2}\log|B|,$$

where $|B| = |K| \cdot |K^{-1}+W|$. Let's apply our developments to this formula. Remembering that

$$K^{-1} = (G^T G + \sigma^2 \cdot I)^{-1} = \frac{1}{\sigma^2}I - \frac{1}{\sigma^4}G^T(I_m + \frac{1}{\sigma^2}GG^T)^{-1}G,$$

we can get $K^{-1}$ from the first term and apply this to $|B|$.

$$|K^{-1}+W| = |\frac{1}{\sigma^2}I + W - \frac{1}{\sigma^4}G^T(I_m + \frac{1}{\sigma^2}GG^T)^{-1}G|,$$

which we can improve by using generalized Weinstein–Aronszajn identity:

$$|A+UCV| = |A| \cdot |C| \cdot |C^{-1} + VA^{-1}U|.$$

$$|\frac{1}{\sigma^2}I_n + W - \frac{1}{\sigma^4}G^T(I_m + \frac{1}{\sigma^2}GG^T)^{-1}G| = |W + \frac{1}{\sigma^2}I_n| \cdot |(I_m + \frac{1}{\sigma^2}GG^T)^{-1}| \cdot |I_m + \frac{1}{\sigma^2}GG^T - \frac{1}{\sigma^4}G(W + \frac{1}{\sigma^2}I_n)^{-1}G^T|$$

$$= \frac{|W + \frac{1}{\sigma^2}I_n|}{|I_m + \frac{1}{\sigma^2}GG^T|} \cdot |I_m + \frac{1}{\sigma^2}GG^T - \frac{1}{\sigma^4}G(W + \frac{1}{\sigma^2}I_n)^{-1}G^T|.$$

Let's polish that with factoring out $\frac{1}{\sigma^2}$ where possible:

$$\frac{\frac{1}{\sigma^{2n}}|\sigma^2 W + I_n|}{\frac{1}{\sigma^{2m}}|\sigma^2 I_m + GG^T|} \cdot \frac{1}{\sigma^{2m}}|\sigma^2 I_m + GG^T - \frac{1}{\sigma^2}G(W + \frac{1}{\sigma^2}I_n)^{-1}G^T|$$

$$= \frac{|\sigma^2 W + I_n|}{\sigma^{2n}|\sigma^2 I_m + GG^T|} \cdot |\sigma^2 I_m + GG^T - G(\sigma^2 W + I_n)^{-1} G^T|.$$

Complexity of this calculation is $O(n + m^3)$, as determinant and inversion of n x n matrices use diagonal matrices only.
Hence, the total marginal log likelihood is:

$$\log(y|X,\theta) = -\frac{1}{2\sigma^2}\hat{f}^T(I_n - G^T(\sigma^2 I_m + GG^T)^{-1}G)\hat{f} + \log p(y|\hat{f})$$

$$-\frac{1}{2}((n-m)\cdot\log\sigma^2 + \log|\sigma^2 I_m + GG^T|)$$

$$-\frac{1}{2}\log\frac{|\sigma^2 W + I_n|}{|\sigma^2 I_m + GG^T|} \cdot |\sigma^2 I_m + GG^T - G(\sigma^2 W + I_n)^{-1} G^T| + \frac{n}{2}\log\sigma^2.$$

One last effort to get rid off $\sigma^{2n}$ in demonimator:

$$\log(y|X,\theta) = -\frac{1}{2\sigma^2}\hat{f}^T(I_n - G^T(\sigma^2 I_m + GG^T)^{-1}G)\hat{f} + \log p(y|\hat{f})$$

$$+\frac{1}{2}(m\cdot\log\sigma^2 - \log|\sigma^2 I_m + GG^T|)$$

$$-\frac{1}{2}\log\frac{|\sigma^2 W + I_n|}{|\sigma^2 I_m + GG^T|} \cdot |\sigma^2 I_m + GG^T - G(\sigma^2 W + I_n)^{-1} G^T|.$$

And summing in the last term:

$$\log(y|X,\theta) = -\frac{1}{2\sigma^2}\hat{f}^T(I_n - G^T(\sigma^2 I_m + GG^T)^{-1}G)\hat{f} + \log p(y|\hat{f})$$

$$+\frac{1}{2}m\cdot\log\sigma^2 - \frac{1}{2}\log|\sigma^2 W + I_n| - \frac{1}{2}\log|\sigma^2 I_m + GG^T - G(\sigma^2 W + I_n)^{-1} G^T|.$$

# 4 Predictions

There are several ways to predicts probabilities. First, using intractable integral. Second, applying squashing function to latent variable. And third, applying squashing function to latent variable by their variance:

$$std - norm - cdf(\frac{mean}{\sqrt{1 + covar}}).$$

# 5 Laplace inference with QR

Next sections are irrelevant to the current work, but it is more convient to have them at hand. Here we develop the same approach but with another matrix decomposition. It was required for some toy experiments. Let suppose noise in covariance matrix again:

$$K = K + \sigma^2 \cdot I = QR + \sigma^2 \cdot I.$$

Then inverse matrix will be:

$$K^{-1} = (QR + \sigma^2 \cdot I)^{-1} = \frac{1}{\sigma^2} I - \frac{1}{\sigma^4} Q(I_m + \frac{1}{\sigma^2} RQ)^{-1} R,$$

$$(K^{-1} + W)^{-1} = ((\frac{1}{\sigma^2} I + W) - \frac{1}{\sigma^4} Q(I_m + \frac{1}{\sigma^2} RQ)^{-1} R)^{-1}.$$

Using matrix inversion lemma one more time and assuming

$$W^*_{inv} = (\frac{1}{\sigma^2} I + W)^{-1},$$

we get final formula:

$$(K^{-1} + W)^{-1} = W^*_{inv} + \frac{1}{\sigma^4} W^*_{inv} Q(I_m + \frac{1}{\sigma^2} RQ - \frac{1}{\sigma^4} RW^*_{inv} Q)^{-1} RW^*_{inv}.$$

# 6 Let QR = K - var I

The reader can notice some collision in designation: $K$ and $K + \sigma^2 I$ denoted by the same symbol $K$. Here we bring more confusion). Let suppose noise in covariance matrix $K^*$:

$$K^* = K + \sigma^2 \cdot I = QR + \sigma^2 \cdot I,$$

where $K = QR$. Then inverse matrix will be:

$$K^{-1} = (QR - \sigma^2 \cdot I + \sigma^2 \cdot I)^{-1} = \frac{1}{\sigma^2} I - \frac{1}{\sigma^4} Q^*(I_m + \frac{1}{\sigma^2} R^* Q^*)^{-1} R^*,$$

where $Q^* R^* = K - \sigma^2 \cdot I$. Then we can estimate the second derivative of posterior as:

$$(K^{-1} + W)^{-1} = ((\frac{1}{\sigma^2} I + W) - \frac{1}{\sigma^4} Q^*(I_m + \frac{1}{\sigma^2} R^* Q^*)^{-1} R^*)^{-1}.$$

Using matrix inversion lemma one more time and assuming

$$W^*_{inv} = (\frac{1}{\sigma^2} I + W)^{-1},$$

we get final formula:

$$(K^{-1} + W)^{-1} = W^*_{inv} + \frac{1}{\sigma^4} W^*_{inv} Q^*(I_m + \frac{1}{\sigma^2} R^* Q^* - \frac{1}{\sigma^4} R^* W^*_{inv} Q^*)^{-1} R^* W^*_{inv}.$$

This approach should be more accurate in terms of estimation of K. However, in our work we suppose that we given $QR = K$. Substracting noise from covarince matrix looks illicit in this case.