Derivatives of lower bound

Michalis K. Titsias
School of Computer Science,
University of Manchester, UK
mtitsias@cs.man.ac.uk

Abstract

1 Useful matrix derivatives

$$\frac{\partial (XY)}{\partial \theta} = X \frac{\partial Y}{\partial \theta} + \frac{\partial X}{\partial \theta} Y \tag{1}$$

$$\frac{\partial K^{-1}}{\partial \theta} = -K^{-1} \frac{\partial K}{\partial \theta} K^{-1} \tag{2}$$

$$\frac{\partial \log |K|}{\partial \theta} = \text{Tr} \left(K^{-1} \frac{\partial K}{\partial \theta} \right) \tag{3}$$

2 Variational lower bound

It can be written in the form

$$F_V = -\frac{n}{2} \log(2\pi) - \frac{n-m}{2} \log \sigma^2 + \frac{1}{2} \log |K_{mm}| - \frac{1}{2} \log |\sigma^2 K_{mm} + K_{mn} K_{nm}| - \frac{1}{2\sigma^2} \mathbf{y}^T \mathbf{y} \\ + \frac{1}{2\sigma^2} \mathbf{y}^T K_{nm} (\sigma^2 K_{mm} + K_{mn} K_{nm})^{-1} K_{mn} \mathbf{y} - \frac{1}{2\sigma^2} \text{tr}(K_{nn}) + - \frac{1}{2\sigma^2} \text{tr}(K_{mm}^{-1} (K_{mn} K_{nm})) \tag{4}$$

We write the above as a sum of the following terms

$$F_0 = -\frac{n}{2} \log(2\pi) - \frac{n-m}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \mathbf{y}^T \mathbf{y} \tag{5}$$

$$F_1 = \frac{1}{2} \log |K_{mm}| \tag{6}$$

$$F_2 = -\frac{1}{2} \log |\sigma^2 K_{mm} + K_{mn} K_{nm}| \tag{7}$$

$$F_3 = \frac{1}{2\sigma^2} \mathbf{y}^T K_{nm} (\sigma^2 K_{mm} + K_{mn} K_{nm})^{-1} K_{mn} \mathbf{y} \tag{8}$$

$$F_4 = -\frac{1}{2\sigma^2} \text{tr}(K_{nn}) \tag{9}$$

$$F_5 = \frac{1}{2\sigma^2} \text{tr}(K_{mm}^{-1} (K_{mn} K_{nm}))$$

3 Derivatives

In the following derivations we make heavily use of the following property of the trace of matrix. In particular, if there a symmetric (implies also square) matrix $\mathcal{A}$ and a square (of same size as $\mathcal{A}$) but possibly not symmetric matrix $\mathcal{B}$, then it holds

$$\text{tr}(\mathcal{A}\mathcal{B}) = \text{tr}(\mathcal{A}\mathcal{B}^T) = \text{tr}(\mathcal{B}^T\mathcal{A}).$$

The proof is obvious since $\text{tr}(\mathcal{A}\mathcal{B}) = \text{tr}(\mathcal{B}\mathcal{A}) = \text{tr}\left((\mathcal{B}\mathcal{A})^T\right) = \text{tr}(\mathcal{A}^T\mathcal{B}^T) = \text{tr}(\mathcal{A}\mathcal{B}^T)$.

$$\frac{\partial F_1}{\partial \boldsymbol{\theta}} = \frac{\partial \log|K_{mm}|}{\partial \theta} = \frac{1}{2}\text{tr}\left(K_{mm}^{-1}\frac{\partial K_{mm}}{\partial \theta}\right) = \frac{1}{2}\text{tr}\left(\frac{\partial K_{mm}}{\partial \theta}K_{mm}^{-1}\right) \tag{10}$$

$$\frac{\partial F_2}{\partial \boldsymbol{\theta}} = -\frac{1}{2}\text{tr}\left(\frac{\partial A}{\partial \boldsymbol{\theta}}A^{-1}\right)$$

where

$$\frac{\partial A}{\partial \theta} = \sigma^2\frac{\partial K_{mm}}{\partial \theta} + \frac{\partial K_{mn}}{\partial \boldsymbol{\theta}}K_{nm} + K_{mn}\frac{\partial K_{nm}}{\partial \theta} = \sigma^2\frac{\partial K_{mm}}{\partial \theta} + \left(K_{mn}\frac{\partial K_{nm}}{\partial \theta}\right)^T + K_{mn}\frac{\partial K_{nm}}{\partial \theta}$$

By substituting the above exression for $\frac{\partial A}{\partial \theta}$, the derivative $\frac{\partial F_2}{\partial \theta}$ is written

$$\frac{\partial F_2}{\partial \theta} = -\frac{\sigma^2}{2}\text{tr}\left(\frac{\partial K_{mm}}{\partial \theta}A^{-1}\right) - \text{tr}\left(\frac{\partial K_{nm}}{\partial \theta}A^{-1}K_{mn}\right)$$

where we used the trace property in eq. ??, with symmetrix matrix $\mathcal{A} = A^{-1}$ and $\mathcal{B}^T = K_{mn}\frac{\partial K_{nm}}{\partial \theta}$ To exress the derivatives for the term F_3, we write first more covneniently in trace form

$$F_3 = \frac{1}{2\sigma^2}\text{tr}\left(K_{nm}A^{-1}K_{mn}\mathbf{y}\mathbf{y}^T\right) = \frac{1}{2\sigma^2}\text{tr}\left(A^{-1}K_{mn}\mathbf{y}\mathbf{y}^TK_{nm}\right)$$

$$\begin{aligned} \frac{\partial F_3}{\partial \boldsymbol{\theta}} &= \frac{1}{2\sigma^2}\text{tr}\left(\frac{\partial A^{-1}}{\partial \theta}K_{mn}\mathbf{y}\mathbf{y}^TK_{nm} + A^{-1}\frac{\partial K_{mn}}{\partial \theta}\mathbf{y}\mathbf{y}^TK_{nm} + A^{-1}K_{mn}\mathbf{y}\mathbf{y}^T\frac{\partial K_{nm}}{\partial \theta}\right) \\ &= \frac{1}{2\sigma^2}\text{tr}\left(\frac{\partial A^{-1}}{\partial \theta}K_{mn}\mathbf{y}\mathbf{y}^TK_{nm}\right) + \frac{1}{\sigma^2}\text{tr}\left(\frac{\partial K_{nm}}{\partial \theta}A^{-1}K_{mn}\mathbf{y}\mathbf{y}^T\right) \end{aligned} \tag{11}$$

where again we took advantage of the symmetry of A^{-1} and apply the property in eq. ?? to simplify the expression. Now by using the fact that $\frac{\partial A^{-1}}{\partial \theta} = -A^{-1}\frac{\partial A}{\partial \theta}A^{-1}$, we have

$$\begin{aligned} \frac{\partial F_3}{\partial \boldsymbol{\theta}} &= -\frac{1}{2\sigma^2}\text{tr}\left(A^{-1}\frac{\partial A}{\partial \theta}A^{-1}K_{mn}\mathbf{y}\mathbf{y}^TK_{nm}\right) + \frac{1}{\sigma^2}\text{tr}\left(\frac{\partial K_{nm}}{\partial \theta}A^{-1}K_{mn}\mathbf{y}\mathbf{y}^T\right) \\ &= -\frac{1}{2\sigma^2}\text{tr}\left(\frac{\partial A}{\partial \theta}A^{-1}K_{mn}\mathbf{y}\mathbf{y}^TK_{nm}A^{-1}\right) + \frac{1}{\sigma^2}\text{tr}\left(\frac{\partial K_{nm}}{\partial \theta}A^{-1}K_{mn}\mathbf{y}\mathbf{y}^T\right) \end{aligned} \tag{12}$$

By using now the $\frac{\partial A}{\partial \theta}$ is given by eq. ??, we further simplify this

$$\frac{\partial F_3}{\partial \boldsymbol{\theta}} = -\frac{1}{2}\text{tr}\left(\frac{\partial K_{mm}}{\partial \theta}A^{-1}K_{mn}\mathbf{y}\mathbf{y}^TK_{nm}A^{-1}\right) - \frac{1}{\sigma^2}\text{tr}\left(\frac{\partial K_{nm}}{\partial \theta}A^{-1}K_{mn}\mathbf{y}\mathbf{y}^TK_{nm}A^{-1}K_{nm}\right) + \frac{1}{\sigma^2}\text{tr}\left(\frac{\partial K_{nm}}{\partial \theta}A^{-1}K_{mn}\mathbf{y}\mathbf{y}^T\right)$$

where again we used the trcae property in eq. ?? by taking advanage now the symmetry of $A^{-1}K_{mn}\mathbf{y}\mathbf{y}^TK_{nm}A^{-1}$

$$\frac{\partial F_4}{\partial \boldsymbol{\theta}} = -\frac{1}{2\sigma^2}\text{tr}\left(K_{nn}\right)$$

$$
\begin{aligned}
\frac{\partial F_5}{\partial \boldsymbol{\theta}} &= \frac{1}{2\sigma^2}\mathrm{tr}\left(\frac{\partial K_{mm}^{-1}}{\partial \boldsymbol{\theta}} K_{mn}K_{nm} + K_{mm}^{-1}\frac{\partial K_{mn}}{\partial \boldsymbol{\theta}} K_{nm} + K_{mm}^{-1}K_{mn}\frac{\partial K_{nm}}{\partial \boldsymbol{\theta}}\right) && (13)\\
&= \frac{1}{2\sigma^2}\mathrm{tr}\left(-K_{mm}^{-1}\frac{\partial K_{mm}}{\partial \boldsymbol{\theta}} K_{mm}^{-1}K_{mn}K_{nm}\right) + \frac{1}{\sigma^2}\mathrm{tr}\left(\frac{\partial K_{nm}}{\partial \boldsymbol{\theta}} K_{mm}^{-1}K_{mn}\right) && (14)\\
&= -\frac{1}{2\sigma^2}\mathrm{tr}\left(\frac{\partial K_{mm}}{\partial \boldsymbol{\theta}} K_{mm}^{-1}K_{mn}K_{nm}K_{mm}^{-1}\right) + \frac{1}{\sigma^2}\mathrm{tr}\left(\frac{\partial K_{nm}}{\partial \boldsymbol{\theta}} K_{mm}^{-1}K_{mn}\right) && (15)
\end{aligned}
$$

where again we used the trace property in eq. ??/ by taking dvantage the symmetry of K_{mm}^{-1}.

3.1 Efficient computation of the derivatives

To exploit now the similarities of the above derivatives so that to discover a effciently ordering of the actual compuations required we write the final forms of the above derivatives and give names to the different terms:

$$
\frac{\partial F_1}{\partial \boldsymbol{\theta}} = \underbrace{\frac{1}{2}\mathrm{tr}\left(\frac{\partial K_{mm}}{\partial \theta} K_{mm}^{-1}\right)}_{(1)}
$$

$$
\frac{\partial F_2}{\partial \theta} = \underbrace{-\frac{\sigma^2}{2}\mathrm{tr}\left(\frac{\partial K_{mm}}{\partial \theta} A^{-1}\right)}_{(2)} \underbrace{-\mathrm{tr}\left(\frac{\partial K_{nm}}{\partial \theta} A^{-1}K_{mn}\right)}_{(3)}
$$

$$
\frac{\partial F_3}{\partial \theta} = \underbrace{-\frac{1}{2}\mathrm{tr}\left(\frac{\partial K_{mm}}{\partial \theta} A^{-1}K_{mn}\mathbf{y}\mathbf{y}^T K_{nm}A^{-1}\right)}_{(4)} \underbrace{-\frac{1}{\sigma^2}\mathrm{tr}\left(\frac{\partial K_{nm}}{\partial \theta} A^{-1}K_{mn}\mathbf{y}\mathbf{y}^T K_{nm}A^{-1}K_{mn}\right)}_{(5)} + \underbrace{\frac{1}{\sigma^2}\mathrm{tr}\left(\frac{\partial K_{nm}}{\partial \theta} A^{-1}K_{mn}\mathbf{y}\mathbf{y}^T\right)}_{(6)}
$$

$$
\frac{\partial F_5}{\partial \theta} = \underbrace{-\frac{1}{2\sigma^2}\mathrm{tr}\left(\frac{\partial K_{mm}}{\partial \theta} K_{mm}^{-1}K_{mn}K_{nm}K_{mm}^{-1}\right)}_{(7)} + \underbrace{\frac{1}{\sigma^2}\mathrm{tr}\left(\frac{\partial K_{nm}}{\partial \boldsymbol{\theta}} K_{mm}^{-1}K_{mn}\right)}_{(8)}
$$

where the blue terms are similar since all have the form $\mathrm{tr}(\frac{\partial K_{mm}}{\partial \boldsymbol{\theta}}\mathcal{C})$ where $\mathcal{C}$ is some (symmetric) matrix os size $m \times m$. Also the red terms are similar since there are all written as $\mathrm{tr}(\frac{\partial K_{m}}{\partial \boldsymbol{\theta}}\mathcal{D})$ where $\mathcal{D}$ is an $m \times n$ matrix. Therefore, we can group the blue and red terms as follows:

$$
(1)+(2)+(4)+(7) = \frac{\sigma^2}{2}\mathrm{tr}\left[\frac{\partial K_{mm}}{\partial \theta}\left(\frac{K_{mm}^{-1}}{\sigma^2} - A^{-1} - \frac{A^{-1}K_{mn}\mathbf{y}}{\sigma}\frac{\mathbf{y}^T K_{nm}A^{-1}}{\sigma} - \frac{K_{mm}^{-1}}{\sigma^2}K_{mn}K_{nm}\frac{K_{mm}^{-1}}{\sigma^2}\right)\right]
$$

$$
(3)+(5)+(6)+(8) = \mathrm{tr}\left[\frac{\partial K_{nm}}{\partial \theta}\left(\left(\frac{K_{mm}^{-1}}{\sigma^2} - A^{-1} - \frac{A^{-1}K_{mn}\mathbf{y}}{\sigma}\frac{\mathbf{y}^T K_{nm}A^{-1}}{\sigma}\right)K_{mn} + \frac{A^{-1}K_{mn}\mathbf{y}}{\sigma^2}\mathbf{y}^T\right)\right]
$$

Impprtantly this shows the the expensive computation $\left(\frac{K_{mm}^{-1}}{\sigma^2} - A^{-1} - \frac{A^{-1}K_{mn}\mathbf{y}}{\sigma}\frac{\mathbf{y}^T K_{nm}A^{-1}}{\sigma}\right)K_{mn}$ between a $m \times m$ and $m \times n$ matrix needs to be compuated before any computation of the derivatives starts. In factr the matrices $\mathcal{C}$ and $\mathcal{D}$ that multiplied to the matrices $\frac{\partial K_{mm}}{\partial \boldsymbol{\theta}}$ and $\frac{\partial K_{nm}}{\partial \boldsymbol{\theta}}$, resepctively, can be precomputed since there are common for all the derivatives with respect to any θ associated with kernel hyperparameter or inducing variable parameter.