Exercise Sheet 6

Robin Schmidt
Probabilistic Inference & Learning

November 24, 2018

Generalized Linear Models

1. Newton Optimization

(a)

Proof. First we recall the taylor expansion in vector notation:

$$\begin{aligned} f(x+\delta x) &= f(x) + \sum_{j=1}^{N} \frac{\partial f(x)}{\partial x_j}\delta x_j + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\partial^2 f(x)}{\partial x_i \partial x_j}\delta x_i \delta x_j + ... \\ f(x+\delta x) &= f(x) + \delta^T x \nabla f(x) + \frac{1}{2}\delta^T x H \delta x + ... \end{aligned}$$

Using the notation stated in the exercise and constructing a quadratic approximation we get:

$$\begin{gathered} L(f_0+\delta) = \widetilde{L}(f_0+\delta) = L(f_0) + \delta^T \nabla L(f_0) + \frac{1}{2}\delta^T B(f_0)\delta \\ f_1 = f_0 + \delta \quad \text{which leads to:} \quad \delta = f_1 - f_0 \end{gathered}$$

Taking the derivative in respect to δ, setting it equal to zero and substituting $\delta = f_1 - f_0$ gives the following:

$$\begin{aligned} 0 \overset{!}{=} \frac{d}{d\delta}\widetilde{L}(f_0+\delta) &= \nabla L(f_0) + B(f_0)\delta \\ -B(f_0)(f_1 - f_0) &= \nabla L(f_0) \\ B(f_0)(f_0 - f_1) &= \nabla L(f_0) \\ f_0 - f_1 &= B^{-1}(f_0)\nabla L(f_0) \\ f_1 &= f_0 - B^{-1}(f_0)\nabla L(f_0) \end{aligned}$$

□

Which shows that the minimum of this approximation lies at $f_1 = f_0 - B^{-1}(f_0)\nabla L(f_0)$ which was to be shown.

(b)

Rewriting the second-order expansion from above and setting the derivation to zero gives the following:

$$\begin{aligned}
L(w+\epsilon) = \widetilde{L}(w+\epsilon) &= L(\phi^T w) + \epsilon^T \nabla L(\phi^T w) + \frac{1}{2}\epsilon^T B(\phi^T w)\epsilon \\
0 &\stackrel{!}{=} \frac{d}{d\epsilon}\widetilde{L}(w+\epsilon) = \nabla L(\phi^T w) + B(\phi^T w)\epsilon \\
-B(\phi^T w)(w - w_0) &= \nabla L(\phi^T w) \\
B(\phi^T w)(w_0 - w) &= \nabla L(\phi^T w) \\
w_0 - w &= B^{-1}(\phi^T w)\nabla L(\phi^T w) \\
w &= w_0 - B^{-1}(\phi^T w)\nabla L(\phi^T w)
\end{aligned}$$

This leads to the Newton step in w with $w = w_0 - B^{-1}(\phi^T w)\nabla L(\phi^T w)$, which was asked in the exercise.

2. Logistic Link Function

(a)

Proof. First we start off by computing $\sigma(y \cdot f)$, taking the logarithm and getting the derivation in respect to f:

$$\begin{aligned}
\sigma(y \cdot f) &= \frac{1}{1 + exp(-yf)} \\
log\ \sigma(y \cdot f) &= log(\frac{1}{1 + exp(-yf)}) \\
\frac{\partial\ log\ \sigma(yf)}{\partial f} &= \frac{\partial}{\partial f} log(\frac{1}{1 + exp(-yf)}) \\
&= \frac{y}{exp(yf) + 1}
\end{aligned}$$

If we only consider the binary classification ($y \in \{-1; 1\}$) and keeping the property of the sigmoid function ($\sigma(x) = 1 - \sigma(-x)$) in mind, we get two cases. One for $y = 1$ and one for $y = -1$. These two cases lead to:

$$\begin{aligned}
\frac{1}{e^f + 1} &= \sigma(-f) = 1 - \sigma(f) \\
\frac{-1}{e^{-f} + 1} &= -\frac{1}{e^{-f} + 1} = -\sigma(f)
\end{aligned}$$

Since both cases include $-\sigma(f)$ and only the first part of the equation is dependent on whether $y = 1$ or $y = -1$, one can easily see that:

$$\frac{\partial\ log\ \sigma(yf)}{\partial f} = \frac{y+1}{2} - \sigma(f)$$

Since $\frac{y+1}{2}$ is equal to 0 for $y = -1$ and equal to 1 for $y = 1$. $\square$

(b)

Proof. Using a similar approach as in (a) we can first take the needed derivative:

$$\begin{aligned}
\frac{\partial^2 log\ \sigma(y \cdot f)}{(\partial f)^2} &= -\frac{y^2 exp(yf)}{(exp(yf)+1)^2} \\
&= \frac{-y^2}{exp(yf)+1} \cdot \frac{exp(yf)}{exp(yf)+1}
\end{aligned}$$

Using the properties stated in (a) we get two cases. For the first case ($y = 1$) we get:

$$\begin{aligned}
&= -\frac{1^2}{e^f+1} \cdot \frac{e^f}{e^f+1} \\
&= -\sigma(-f) \cdot \frac{e^f}{e^f+1} \\
&= (\sigma(f)-1) \cdot \frac{e^f}{e^f+1} \\
&= (\sigma(f)-1) \cdot (e^f * \sigma(-f)) \\
&= (\sigma(f)-1) \cdot (e^f * (1-\sigma(f))) \\
&= (\sigma(f)-1) \cdot (e^f * (1-\frac{1}{exp(-f)+1})) \\
&= (\sigma(f)-1) \cdot (e^f - \frac{e^f}{e^-f+1}) \\
&= (\sigma(f)-1) \cdot (\frac{e^f \cdot (e^{-f}+1) - e^f}{e^{-f}+1}) \\
&= (\sigma(f)-1) \cdot (\frac{e^f \cdot e^{-f} + e^f - e^f}{e^{-f}+1}) \\
&= (\sigma(f)-1) \cdot (\frac{1+0}{e^{-f}+1}) \\
&= (\sigma(f)-1) \cdot (\frac{1}{e^{-f}+1}) \\
&= (\sigma(f)-1) \cdot \sigma(f) \\
&= \sigma(f) \cdot (\sigma(f)-1) \\
&= -\sigma(f) \cdot (1-\sigma(f))
\end{aligned}$$

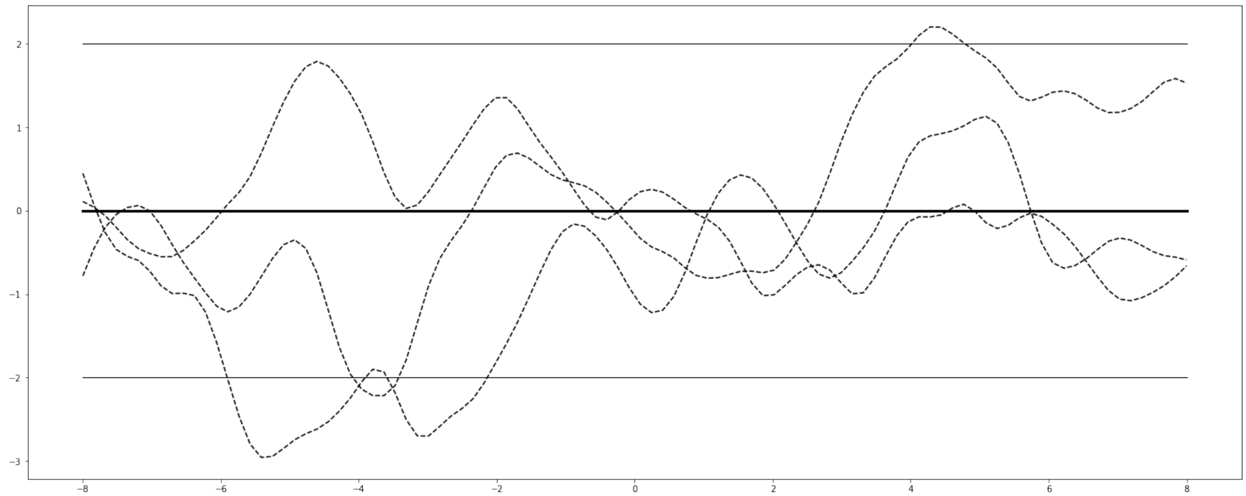
For the second case ($y = -1$) we get:

$$
\begin{aligned}
&= -\frac{(-1)^2}{e^{-f}+1} \cdot \frac{e^{-f}}{e^{-f}+1} \\
&= -\sigma(f) \cdot \frac{e^{-f}}{e^{-f}+1} \\
&= -\sigma(f) \cdot \frac{e^{-f}}{e^{-f}+1} \\
&= -\sigma(f) \cdot (e^{-f} * \sigma(f)) \\
&= -\sigma(f) \cdot (e^{-f} * (1 - \sigma(-f))) \\
&= -\sigma(f) \cdot (e^{-f} * (1 - \frac{1}{e^{f}+1})) \\
&= -\sigma(f) \cdot (e^{-f} - \frac{e^{-f}}{e^{f}+1}) \\
&= -\sigma(f) \cdot (\frac{e^{-f} \cdot (e^{f}+1) - e^{-f}}{e^{f}+1}) \\
&= -\sigma(f) \cdot (\frac{e^{-f} \cdot e^{f} + e^{-f} - e^{-f}}{e^{f}+1}) \\
&= -\sigma(f) \cdot (\frac{1}{e^{f}+1}) \\
&= -\sigma(f) \cdot \sigma(-f) \\
&= -\sigma(f) \cdot (1 - \sigma(f))
\end{aligned}
$$

As we can see both cases have the same result, which is $-\sigma(f) \cdot (1 - \sigma(f))$. This was to be shown. $\square$

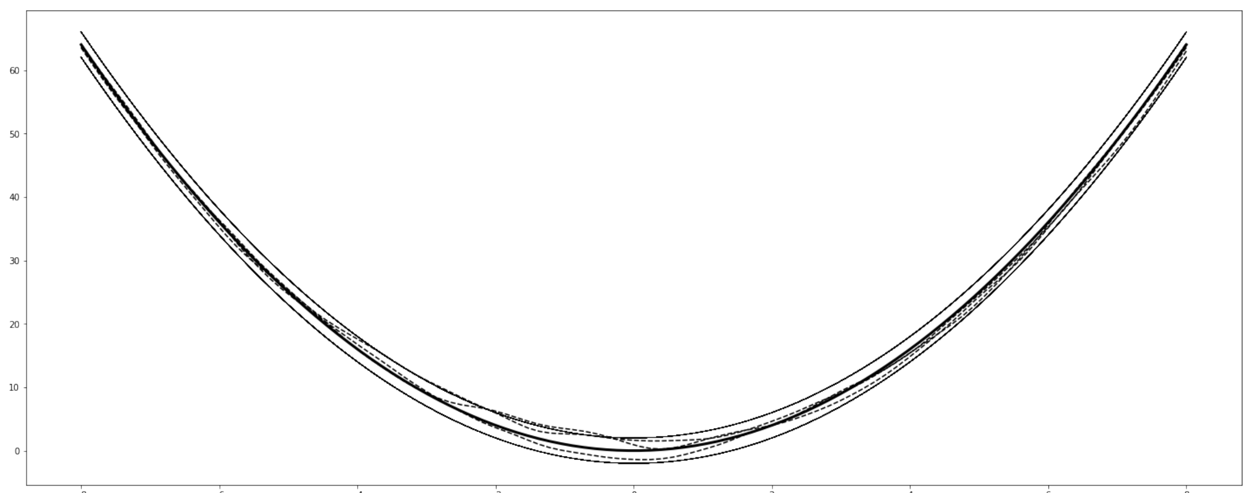
Basic Properties of Gaussian Processes

(a)



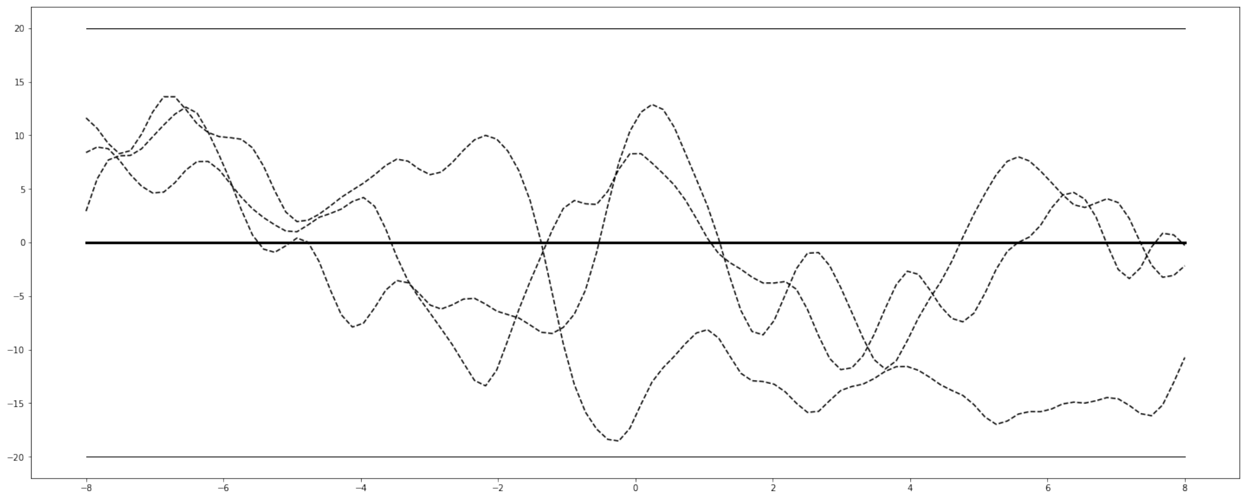
please see "Schmidt_Robin_Ex06_3a.ipynb" for code.

(b)



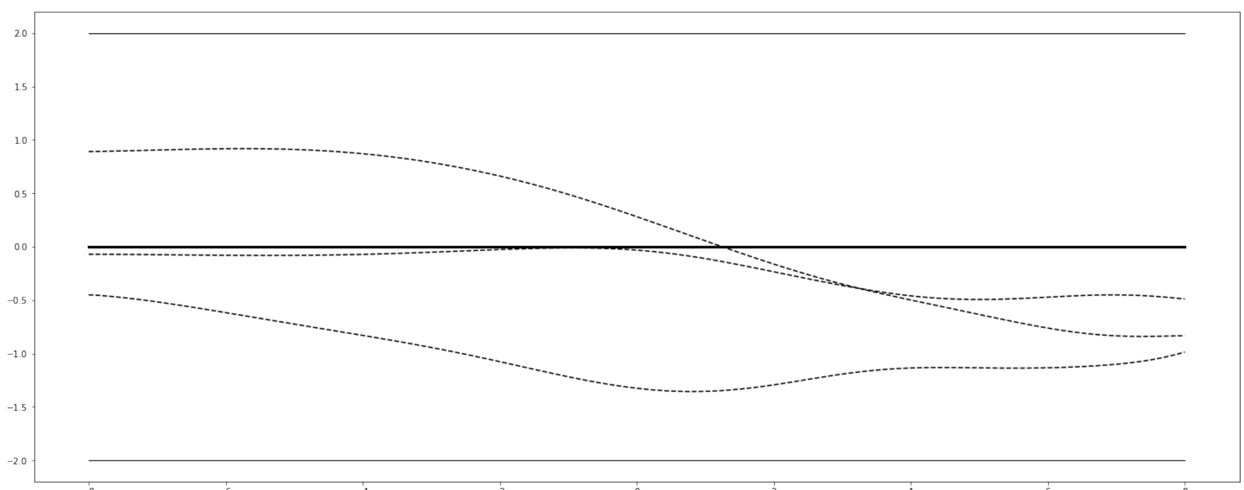
please see "Schmidt_Robin_Ex06_3b.ipynb" for code.

(c)



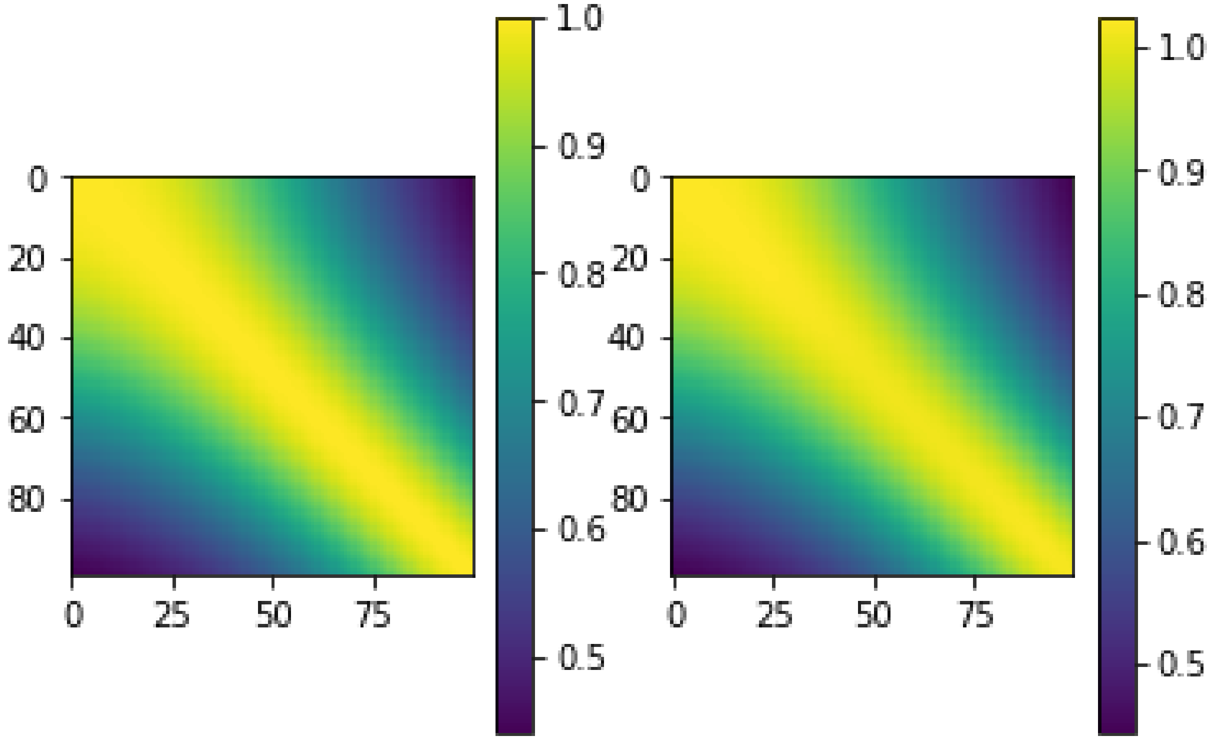
please see "Schmidt_Robin_Ex06_3c.ipynb" for code.

(d)



please see "Schmidt_Robin_Ex06_3d.ipynb" for code.

(e)



please see "Schmidt_Robin_Ex06_3e.ipynb" for code.

They look similar due to the "law of large numbers", which describes the result of performing the same experiment a large number of times. According to the law, the average of the results obtained from a large number of trials should be close to the expected value, and will tend to become closer as more trials are performed.