Theory Question

a) $$p(f_X \mid y, X) = \frac{p(y \mid f_X, X)\, p(f_X \mid X)}{p(y \mid X)}$$, using Bayes' Rule

$$p(f \mid y) = \frac{p(y \mid f)\, p(f)}{p(y)}$$ → Without loss of generality for continuous X

$$\log(p(f_X \mid y, X)) = \log p(y \mid f_X) + \log p(f_X) - \log p(y)$$

$$= \log \prod_{i=1}^{N} \sigma(y_i f(x_i)) + \log p(f_X) + \text{constant term}$$

$$p(f_X) = \mathcal{N}(f_X ; \underbrace{m(X)}_{m_X}, \underbrace{k(X,X)}_{K_{XX}})$$

$$\log(p(f_X \mid y, X)) \cong \left(-\sum_{i=1}^{N} \log(1 + \exp(-y_i f(x_i)))\right) - \frac{1}{2}(f_X - m_X)^T K_{XX}^{-1} (f_X - m_X)$$

b) $\nabla_{f_x} \log(p(f_x \mid y, X)) = \sum_{i=1}^{N} \nabla_{f_x} \log(\sigma(y_i, f(x_i))) - K_{xx}^{-1}(f_x - m_x)$

$= \underbrace{\sum_{i=1}^{N} -\nabla f_x \log(1+\exp(-y_i f x_i))} - K_{xx}^{-1}(f_x - m_x)$

where $-\dfrac{\partial(\log(1+\exp(-y_i f x_i))}{\partial f x_j} = \delta_{ij}\left(\dfrac{y_i+1}{2} - \sigma(f x_i)\right)$

where $y_i \in \{-1, 1\}$

$\Rightarrow \dfrac{y_i+1}{2} \in \{0, 1\}$

Note that $-\dfrac{\partial}{\partial f x_i} \log(1+\exp(-y_i f x_i)) = \dfrac{y_i \exp(-y_i f x_i)}{1+\exp(-y_i f x_i)}$

① $y_i = -1 \Rightarrow -\dfrac{\exp(f x_i)}{1+\exp(f x_i)} = \dfrac{-1}{1+\exp(-f x_i)} = -\sigma(f x_i)$

② $y_i = 1 \Rightarrow \dfrac{\exp(-f x_i)}{1+\exp(-f x_i)} = 1 - \dfrac{1}{1+\exp(-f x_i)} = 1 - \sigma(f x_i)$

① ② $\Rightarrow \dfrac{y_i+1}{2} - \sigma(f x_i)$

$$\nabla_{f_x} \log(p(f_x \mid y)) = \begin{bmatrix} \frac{y_1+1}{2} - \sigma(f(x_1)) \\ \frac{y_2+1}{2} - \sigma(f(x_2)) \\ \vdots \\ \frac{y_n+1}{2} - \sigma(f(x_n)) \end{bmatrix} - K_{xx}^{-1}(f_x - m_x)$$

c) We know that at the mode, $f_x = \hat{f}_x$:

$$\nabla_{f_x} \log p(f_x \mid y, x)\Big|_{f_x=\hat{f}_x} = 0$$

Using part b $\rightarrow$ $\nabla_{f_x} \log(p(y \mid f_x)) - k_{xx}^{-1}(f_x - m_x) = 0$, at the mode

⊛ $\nabla \log(p(y \mid \hat{f}_x)) = k_{xx}^{-1}(\hat{f}_x - m_x)$, at the mode

Way 1

$k_{xx} \nabla \log(p(y \mid f_x))$ $\underbrace{k_{xx} k_{xx}^{-1}}_{=1,\ \text{spd}}(f_x - m_x)$, at the mode

$\Rightarrow$ $m_x + k_{xx} \nabla \log(p(y \mid f_x)) = f_x$, at the mode

From the last result, it can be seen that

$$\mathbb{E}_q[f(\cdot)] = m_{\cdot} + k_{\cdot x} \nabla \log(p(y \mid \hat{f}_x))$$

Way 2

More rigorously, as depicted also in lecture 14, slide 17:

where $q(f_x) = \mathcal{N}(f_x; \hat{f}_x, \Sigma)$

$$q(f_x \mid y) = \int p(f_x \mid \vec{f}_x)\, q(\vec{f}_x)\, d\vec{f}_x$$

$$= \mathcal{N}(f_x;\ \underline{m_x + k_{xX} k_{XX}^{-1}(\vec{f}_X - m_X)},\ \Sigma_{\text{posterior}})$$

Thus, $\mathbb{E}_q[f(\cdot)] = m(\cdot) + k_{\cdot x} k_{xx}^{-1}(\hat{f}_x - m_x)$ as expected value is precisely the mean of gaussian.

Using ⊛ to replace $k_{xx}^{-1}(\hat{f}_x - m_x)$ with $\nabla \log(p(y \mid \hat{f}_x))$, the MAP estimate can be written as:

$$\boxed{\mathbb{E}_q[f(\cdot)] = m(\cdot) + k_{\cdot x} \nabla \log(p(y \mid \hat{f}_x))}$$

Q.E.D.

d) Utilizing the results of part b, the explicit form can be written as:

$$\nabla \log(p(y)\hat{f}_x) = \sum_{i=1}^{N} \nabla \hat{f}_x \log(\sigma(y_i \hat{f}(x_i)))$$
$$= \sum_{i=1}^{N} \nabla \hat{f}_x [-\log(1+\exp(-y_i \hat{f}(x_i)))]$$

with $\frac{\partial}{\partial \hat{f}_{x_i}}(-\log(1+\exp(-y_i \hat{f}(x_i)))) = \frac{y_i+1}{2} - \sigma(\hat{f}(x_i))$

and $\frac{\partial \log \sigma(y_i \hat{f}(x_i))}{\partial \hat{f}_{x_j}} = \delta_{ij}\left(\frac{y_i+1}{2} - \sigma(\hat{f}(x_i))\right)$

Observe that when $|\hat{f}(x_i)| \gg 1$, there are two cases: either the $\hat{f}(x_i)$ is too small or too large:

① when $\hat{f}(x_i) \gg 1$, $\sigma(\hat{f}(x_i)) = \frac{1}{1+\exp(-\hat{f}(x_i))} \approx 1$ as $\lim_{x\to\infty} \exp(-x) = 0$

and the gradient term becomes, $\frac{y_i+1}{2} - \sigma(\hat{f}(x_i)) = \frac{1+1}{2} - 1 = 0$

② when $\hat{f}(x_i) \ll 1$, $\sigma(\hat{f}(x_i)) = \frac{1}{1+\exp(-\hat{f}(x_i))} \approx 0$ as $\lim_{x\to-\infty} \exp(-x) = \infty$

and the gradient term becomes, $\frac{y_i+1}{2} - \sigma(\hat{f}(x_i)) = \frac{-1+1}{2} - 0 = 0$

Therefore, training points far from the decision boundary do almost not contribute to this estimate $E_q(f(\cdot))$.

e) We consider the normalization of the suggested likelihood

$\sum_{y_i \in \{-1,1\}} c\, r(y_i; f(x_i))$ with $r(y_i; f(x_i)) = \exp[-\max(0, 1 - y_i f(x_i))]$

Separate negatives and positives

$$\sum_{y_i=-1} c \exp[-\max(0, 1+f(x_i))] + \sum_{y_i=1} c \exp[-\max(0, 1-f(x_i))]$$

without loss of generality, consider for individual training points in the positive and negative classes. Then for two such points, we have

$$\underbrace{c \exp[-\max(0, 1+f(x_i))]}_{y_i=-1 \text{ case}} + \underbrace{c \exp[-\max(0, 1-f(x_i))]}_{y_i=1 \text{ case}}$$

Without loss of generality, consider the case $f(x_i) > 1$, then:

$$c \exp[-(1+f(x_i))] + c \exp[0] = c[1 + e^{-(1+f(x_i))}]$$

The normalization constant, c should include the term:

$$c = c_1 \frac{1}{1+e^{-(1+f(x_i))}}$$ yet this is not independent of $f(x_i)$

Thus $r(y_i, f(x_i))$ is not a family of probability distributions and can't be a likelihood.