Exercise 1: EMR

Task 1:

$$D = \{(x_i, y_i)\}_{i=1}^{N} \qquad \mathcal{L}(y, f(x))$$

$$R_{emp}(f \mid D) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(y_i, f(x_i))$$

$$R(f) = \mathbb{E}_{p(x,y)} \left[\mathcal{L}(\tilde{f}(x) = y, f(x)) \right]$$

To show $R_{emp}(f \mid D) \xrightarrow{a.s.} R(f)$:

Since we have i.i.d random variables, the random variables are drawn from $p(x,y)$, and $N \to \infty$, we can apply the LLN to $R_{emp}(f \mid D)$

$$\Rightarrow \quad R_{emp}(f \mid D) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(y_i, f(x_i))$$

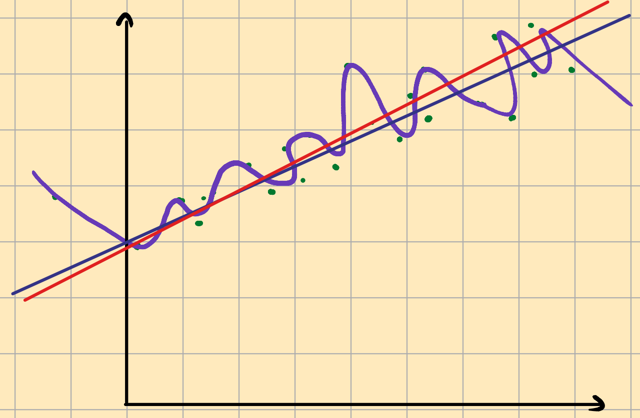
$$= \mathbb{E}_{p(x,y)} \left[\mathcal{L}(y, f(x)) \right] = R(f)$$

Task 2:

a) The expected risk can be decomposed into
$R(f) = Bias^2(f) + Var(f)$

The reason for this can be that the model in $\mathcal{H}_2$ is complex enough to overfit to the data. Meaning on the given dataset the model f_2 fits to all datapoints instead of learning the distribution behind the data. This leads to a high $R(f_2)$ but a low empirical risk $R_{emp}(f \mid D)$. Because the model fits the training data but is not able to generalize to new datapoints
In terms of bias and variance: The bias is getting lower but the variance increases
The f_1 model does not have the complexity to overfit the data: higer bias but lower variance.

b)



c) Since in k-fold cross validation we validate the model on the k'th fold we can see how well the model generalizes to unseen data, penalizing the model with overfitting due to the large validation error. So if the model tends to overfit, we have the chance to detect that and opt for the model with less validation error.

Task 3:

a) $\hat{y}_{map} = \arg\max_k \; p(y = k \mid x)$

$$R(f) = \mathbb{E}_{p(x,y)}\left[\mathbb{1}_{\{f(x) \neq y\}} \right]$$

$$= \mathbb{E}_{p(x)}\left[\mathbb{E}_{p(y \mid x)}\left[\mathbb{1}_{\{f(x) \neq y\}} \mid x \right]\right]$$

$$= \int \mathbb{E}_{p(y \mid x)}\left[\mathbb{1}_{\{f(x) \neq y\}} \mid x \right] \underbrace{p(x)}_{\text{non negative}} \, dx$$

$\Rightarrow$ minimize $\mathbb{E}_{p(y \mid x)}\left[\mathbb{1}_{\{f(x) \neq y\}} \mid x \right]$

$$\mathbb{E}_{p(y \mid x)}\left[\mathbb{1}_{\{f(x) \neq y\}} \mid x \right] = \begin{cases} P(y = 1 \mid x) & \text{if } f(x) = 0 \\ P(y = 0 \mid x) & \text{if } f(x) = 1 \end{cases}$$

$\Rightarrow$ To minimize this choose the larger posterior probability:

if $p(y=1 \mid x) > p(y=0 \mid x)$ then choose $\hat{f}(x) = 1$,
and 0 else.

This is exactly the MAP decision rule.

b)