1,1

$h(\theta): \mathbb{R}^m \to \mathbb{R}$, $\sigma \in \mathbb{R}^m$, $\underline{y} \in \mathbb{R}^n$

$$h(\theta) = \frac{1}{2}\|f(\theta)-y\|^3$$

$$h'(\theta) = \frac{3}{2}\|f(\theta)-y\|^2 \cdot \|f(\theta)-y\|'$$

$$\|f(\theta)-y\|' = \sqrt{\sum_{i=1}^{n}(f_i(\theta)-y_i)^2}\,'$$

$$= \frac{1 \cdot 2(f_i(\theta)-y_i) \cdot f_i'(\theta)}{2\sqrt{\sum_{i=1}^{n}(f_i(\theta)-y_i)^2}}$$

$$h'(\theta) = \frac{\frac{3}{2}\|f(\theta)-y\|^{\cancel{2}} \cdot \cancel{2}(f(\theta)-y) \cdot f'(\theta)}{\cancel{2\|f(\theta)-y\|}}$$

$$= \frac{3}{2}\|f(\theta)-y\|(f(\theta)-y) \cdot f'(\theta)$$

1.2

$$f(q) = \begin{bmatrix} q_1 \cdot q_2 \\ q_1^2 + q_2^2 \\ q_1 \end{bmatrix}$$

$$f_1'(q) = \begin{bmatrix} q_2 \\ 2q_1 \\ 1 \end{bmatrix}, \quad f_2'(q) = \begin{bmatrix} q_1 \\ 2q_2 \\ 0 \end{bmatrix}$$

$$f'(q) = \begin{bmatrix} q_2 & 2q_1 & 1 \\ q_1 & 2q_2 & 0 \end{bmatrix}$$

$$q = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \quad y = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \quad f\left(\begin{bmatrix} 1 \\ 2 \end{bmatrix}\right) = \begin{bmatrix} 2 \\ 5 \\ 0 \end{bmatrix}$$

$$f\left(\begin{bmatrix} 1 \\ 2 \end{bmatrix}\right) - \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 4 \\ 0 \end{bmatrix}$$

$$h'\left(\begin{bmatrix} 1 \\ 2 \end{bmatrix}\right) = \frac{3}{2} \left\| \begin{bmatrix} 1 \\ 4 \\ 0 \end{bmatrix} \right\| \cdot \begin{bmatrix} 1 \\ 4 \\ 0 \end{bmatrix} \cdot \begin{bmatrix} 2 & 2 & 1 \\ 1 & 4 & 0 \end{bmatrix}$$

$$= \frac{3}{2} \cdot \sqrt{17} \cdot \begin{bmatrix} 1 \cdot 2 + 1 \cdot 1 \\ 4 \cdot 2 + 4 \cdot 4 \\ 0 \end{bmatrix} = \frac{3\sqrt{17}}{2} \begin{bmatrix} 3 \\ 24 \\ 0 \end{bmatrix}$$

$$S'(X)_j \in \mathbb{R}^{K}$$

$$S'(X) \in \mathbb{R}^{K\times K}$$

$$S(X)_j = \frac{e^{x_j}}{\sum_{l=1}^{K} e^{x_l}}$$

$$S'(x_i)_j = \left(\frac{e^{x_j}}{\sum_{l=1}^{K} e^{x_l}}\right)' = \frac{(e^{x_j})' \sum_{l=1}^{K} e^{x_l} - (e^{x_i})' e^{x_j}}{\left(\sum_{l=1}^{K} e^{x_l}\right)^2} = \frac{e^{x_j} \cdot \sum_{l=1}^{K} e^{x_l} - e^{x_i} e^{x_j}}{\left(\sum_{l=1}^{K} e^{x_l}\right)^2}$$

$$= \frac{e^{x_j}\left(\sum_{l=1}^{K} e^{x_l} - e^{x_i}\right)}{\left(\sum_{l=1}^{K} e^{x_l}\right)^2} = \frac{\sum_{l \in [K]} e^{x_j} \cdot e^{x_{l \neq i}}}{\sum_{l=1}^{K} e^{x_l} \cdot \sum_{l=1}^{K} e^{x_l}}$$

$$= \frac{e^{x_j}}{e^{x_i} \sum_{l=1}^{K} e^{x_l}} \Longrightarrow S'(X)_j = \begin{bmatrix} \frac{e^{x_j}}{e^{x_1}\sum_{l=1}^{K} e^{x_l}} & \cdots & \frac{e^{x_j}}{e^{x_K}\sum_{l=1}^{K} e^{x_l}} \end{bmatrix}$$

$$S'(X) = \begin{array}{c} S'(X)_1 \\ S'(X)_2 \\ \vdots \\ S'(X)_K \end{array} \begin{bmatrix} \frac{1}{\sum_{l=1}^{K} e^{x_l}} & \cdots & \frac{e^{x_1}}{e^{x_K}\sum_{l=1}^{K} e^{x_l}} \\ \vdots & \ddots & \vdots \\ \frac{e^{x_K}}{e^{x_1}\sum_{l=1}^{K} e^{x_l}} & \cdots & \frac{1}{\sum_{l=1}^{K} e^{x_l}} \end{bmatrix} \in \mathbb{R}^{K\times K}$$

>

Scanerio 1

- The average error over 100 experiments when calculating the error over the test set for the chosen prophet of the experiment is 0.30027
- The number of selected wisely the best prophet is prophet is 64 times, which correlates with the probability to choose the best prophet
- Approximation and estimation errors are presented in the matching columns.

Scanerio 2

- The average error over 100 experiments when calculating the error over the test set for the chosen prophet of the experiment is 0.22464
- The number of selected wisely the best prophet is 84 times.
- This time we chose the better prophet more times because the samples number is higher, gives the worse prophet less chances to correct more than the better prophet, according to their true risks.

Scanerio 3

- The average error over 100 experiments when calculating the error over the test set for the chosen prophet of the experiment is 0.09509
- The number of selected wisely the best prophet is 5.
- The number of prophets chosen that are less than 1% worse from the best is 10 prophets.
- Approximation error is 0.006 over all experiments because the prophets set doesn't change over the experiments, and the approximation error is the true risk of the best prophet of the set.
  (the best prophet from the set doesn't change)
- The estimation error over 100 experiments when calculating the error over the test set for the chosen prophet of the experiment is 0.08732
- If the error rates were uniformly distributed between [0, 0.5] instead of [0, 1], we would expect that more prophets would correct over a training set of 10 games, making us to choose more prophets that their true risks are much higher than their empirical risks. In this manner, the estimation error should rise because its evaluating the distance between the bias and the chosen prophet true risk. However, the best prophet true risk remain the same thus the approximation error wouldn't change.

Scanerio 4

- The average error over 100 experiments when calculating the error over the test set for the chosen prophet of the experiment is 0.00687.
- The number of selected wisely the best prophet is 48.
- The number of prophets chosen that are less than 1% worse from the best prophet is 99 prophets.
- Approximation error is 0.006 over all experiments because the prophets set doesn't change over the experiments, and the approximation error is the true risk of the best prophet of the set.
- The estimation error over 100 experiments when calculating the error over the test set for the chosen prophet of the experiment is 0.080.
- the generalization error of the selected prophet differ from calculating on train vs test set in a way that calculating on the train set will raise a lower generalization error. Because 1000 samples from the train ser used to determine the empirical error of the prophet, so lots sample from train set will contribute 0 to the empirical error (because we chose the ERM).
  If we calculate on the test set, all the samples are 'new' to the prophet in a way that we didn't choose the prophet along these samples, so the generalization gap in this case will probably raise.

Scenario 5 [average error,estimation error,approximation error]

K/N	N=1	N=10	N=50	N=1000
K=2	[0.01785, 0.01718, 0.00884]	[0.00696, 0.00546, 0.00884]	[0.003, 0.00119, 0.00884]	[0.003, 0.00119, 0.00884]
K=5	[0.03961, 0.01679, 0.02298]	[0.02638, 0.00675, 0.02298]	[0.01348, 0.00301, 0.02298]	[0.003, 0.01295, 0.02298]
K=10	[0.08014, 0.08243, 0.00049]	[0.09974, 0.10043, 0.00049]	[0.12188, 0.11319, 0.00049]	[0.117, 0.10299, 0.00049]
K=50	[0.10994, 0.09961, 0.00941]	[0.10994, 0.09961, 0.00941]	[0.07784, 0.06533, 0.00941]	[0.07784, 0.06533, 0.00941]

We can explain the results on the table with the bias-average error over the test set trade-off. In the scenario when k =2, if we look at this row when we take small number of examples(5,10) we chose wrongly the prophet who had the best results on the data but not on the test set, these scenarios are similar to scenarios 1 and 2.

If we take a set of more prophets, k=5 and look at this row we see that the average error on the test set is getting better as N is larger, and estimation error pattern is decreasing as N is larger. We can explain the results when K=5,N=50 and K=5,N=1000 by overfitting of the not best prophet over 50 examples but not over 1000. And indeed the average error over the test set is better when N=1000.

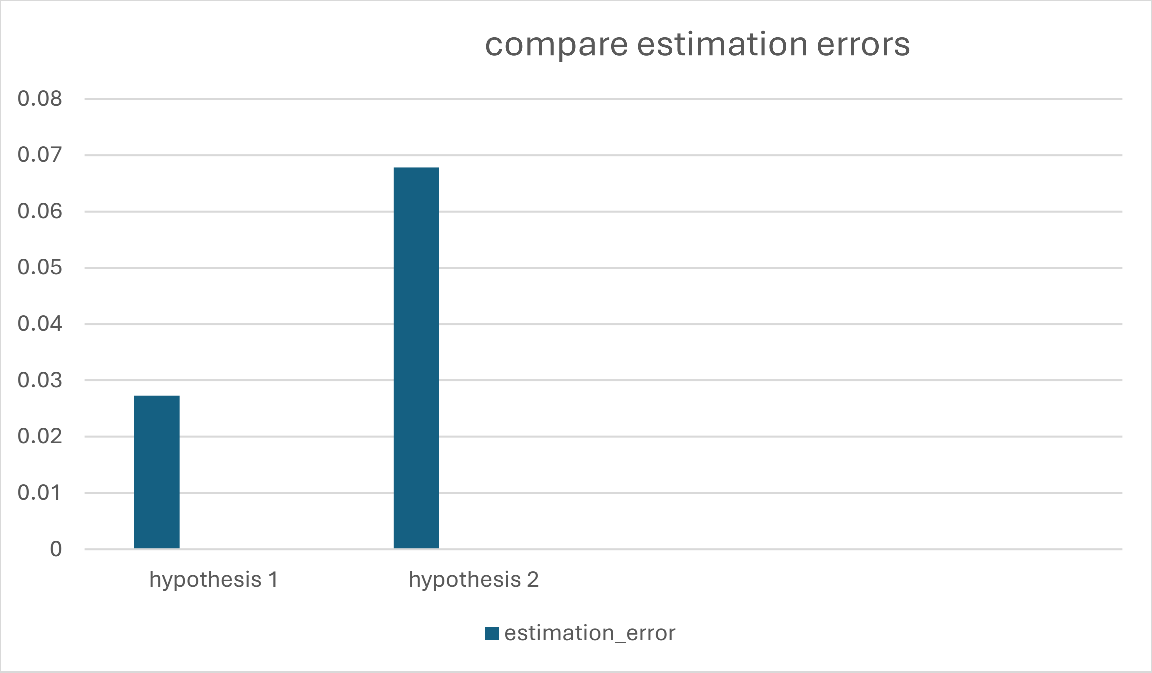
We can include these two rows and compare it to scenarios 1-2 when the number of prophets is low there is less overfitting, meaning that we choose the best prophet from the set and a good prophet over the test set, despite the small set of prophets.

When we look on rows when K=10 and K=50, that are more similar to scenarios 3 and 4, we see that the approximation error is lower and estimation error is higher as N become greater.

We can explain these results once again with the trade-off. When the number of prophets is higher, assume we randomly chose a set of professional prophets, the probability to choose a better prophet as the best raises, comparing to the previous cases (K=2,K=5). But if we look at the estimation error and average error, we an increase pattern from the overfitting of the prophets in the set. When we have more prophets in the set, the probability for one of them to correct over the train set raises, and so he probability to choose the not best prophet raises.

Scenario 6

	Average error	Estimation error	Approximation error
Hypothesis 1	0.3527	0.0273	0.3230
Hypothesis 2	0.3207	0.0678	0.2500



We can see from the experiment results that the average error of both hypothesis are slightly the same as they both around 32%-35%. However, the estimation error in hypothesis 2 is 3 times greater than the estimation error in hypothesis 1. Because the estimation error is set to be the ERM – bias, high estimation error suggest that the chosen prophet is further from the best prophet, makes the generalization gap higher. Another side of the tradeoff complexity is the effect of the number of prophets on the approximation error. We can see in hypotheses 2, which has much more prophet than hypotheses 1, that the bias is the smalles value of the distribution set (0.25), when in hypotheses 1 the bias is almost 10% higher than the smallest value in the set (0.323,03).

To conclude, hypothesis 2 has more prophets, corelates with lower bias and high estimation error, and hypothesis 1 has less prophets, corelates with higher bias and lower estimation error.