Intro to Machine Learning Homework Assignment 5

Goktug Saatcioglu

NetId: gs2417

1. We know that the determinant of a diagonal matrix is given by the multiplication of its diagonal elements. In our case the product

$$|\Sigma| = \prod_{i=1}^{d} \sigma_i^2$$

gives us the determinant of the covariance matrix Σ. Furthermore, the inverse of a diagonal matrix is given as 1 over its diagonal elements. In our case

$$\Sigma^{-1} = \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma_2^2} & \dots & 0 \\ \vdots & 0 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & \frac{1}{\sigma_d^2} \end{bmatrix}$$

gives us the inverse of the covariance matrix Σ, denoted by Σ^{-1}. We begin by re-writing $\frac{1}{Z}$ in product form. $\frac{1}{Z}$ as given in the question is defined as

$$\frac{1}{Z} = (2\pi)^{-\frac{d}{2}} |\Sigma|^{-\frac{1}{2}}$$

which can then be re-written as

$$\frac{1}{Z} = \frac{1}{\sqrt{(2\pi)^d}} \frac{1}{\sqrt{|\Sigma|}}.$$

Let's first consider the second fraction involving the determinant of the covariance matrix. We know that

$$\frac{1}{\sqrt{|\Sigma|}} = \frac{1}{\sqrt{|\Sigma| = \prod_{i=1}^{d} \sigma_i^2}} = \frac{1}{|\Sigma| = \prod_{i=1}^{d} \sigma_i} = \prod_{i=1}^{d} \frac{1}{\sigma_i}.$$

Similarly, we can now re-write the first fraction in product form as follows

$$\frac{1}{\sqrt{(2\pi)^d}} = \frac{1}{\sqrt{\prod_{i=1}^{d} 2\pi}} = \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi}}.$$

Using our results from above, we can re-write $\frac{1}{Z}$ as follows

$$\frac{1}{Z} = \frac{1}{\sqrt{(2\pi)^d}} \frac{1}{\sqrt{|\Sigma|}} = \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi}} \prod_{i=1}^{d} \frac{1}{\sigma_i} = \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi}\sigma_i}.$$

Now let's consider the term

$$\exp(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu))$$

which we then consider first the multiplication of Σ^{-1}, which is a $d \times d$ matrix, by $x - \mu$, which is a $d \times 1$ column matrix. This multiplication gives us

$$\Sigma^{-1}(x-\mu) = \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma_2^2} & \dots & 0 \\ \vdots & 0 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & \frac{1}{\sigma_d^2} \end{bmatrix} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \\ \vdots \\ x_d - \mu_d \end{bmatrix} = \begin{bmatrix} \frac{x_1-\mu_1}{\sigma_1^2} \\ \frac{x_2-\mu_2}{\sigma_2^2} \\ \vdots \\ \frac{x_d-\mu_d}{\sigma_d^2} \end{bmatrix}.$$

Then we multiply the $1 \times d$ row vector by the $d \times 1$ column vector $(x-\mu)^\top$ by $\Sigma^{-1}(x-\mu)$ (which is essentially a dot product) to get

$$(x-\mu)^\top \Sigma^{-1}(x-\mu) = \begin{bmatrix} x_1 - \mu_1 & x_2 - \mu_2 & \dots & x_d - \mu_d \end{bmatrix} \begin{bmatrix} \frac{x_1-\mu_1}{\sigma_1^2} \\ \frac{x_2-\mu_2}{\sigma_2^2} \\ \vdots \\ \frac{x_d-\mu_d}{\sigma_d^2} \end{bmatrix}$$

which then gives us

$$\frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} + \dots + \frac{(x_d-\mu_d)^2}{\sigma_d^2} = \sum_{i=1}^{d} \frac{(x_i-\mu_i)^2}{\sigma_i^2}.$$

Thus, we see that

$$\exp(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)) = \exp(-\frac{1}{2}\sum_{i=1}^{d} \frac{(x_i-\mu_i)^2}{\sigma_i^2}),$$

which then can be written in product form using the property $\exp(x+y) = \exp(x)\exp(y)$ which gives us

$$\exp(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)) = \prod_{i=1}^{d} \exp(-\frac{1}{2}\frac{(x_i-\mu_i)^2}{\sigma_i^2}).$$

Finally, we can combine all our results together to re-write $f(x)$ as follows

$$\begin{aligned} f(x) &= \frac{1}{Z}\exp(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)) \\ &= \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi}\sigma_i} \prod_{i=1}^{d} \exp(-\frac{1}{2}\frac{(x_i-\mu_i)^2}{\sigma_i^2}) \\ &= \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi}\sigma_i} \exp(-\frac{1}{2}\frac{1}{\sigma_i^2}(x_i-\mu_i)^2) \\ \therefore f(x) &= \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi}\sigma_i} \exp(-\frac{1}{2}\frac{1}{\sigma_i^2}(x_i-\mu_i)^2), \end{aligned}$$

which proves the property we seek to prove.

2. (a) Consider the conditional probability of an event X given event Y which is defined as

$$p(X \mid Y) = \frac{p(X \cap Y)}{p(Y)},$$

which then can be re-written as

$$p(X \cap Y) = p(X \mid Y)p(Y).$$

Also consider the conitional probability of an event Y given event X which is defined as

$$p(Y \mid X) = \frac{p(Y \cap X)}{p(X)}$$

which can then be re-written as

$$p(Y \cap X) = p(Y \mid X)p(X).$$

Since $p(X \cap Y) = p(Y \cap X)$, we can set both sides to each other and get the following equation

$$p(X \mid Y)p(Y) = p(Y \mid X)p(X).$$

Dividing both sides by $p(X)$ gives us

$$p(Y \mid X) = \frac{p(X \mid Y)p(Y)}{p(X)},$$

which proves that Bayes' rule is true.

(b) $\mathbb{E}[X+Y]$ where X and Y are discrete random variables is given by

$$\mathbb{E}[X+Y] = \sum_{x,y\in\Omega} (x+y)p(x,y).$$

We also know that from the law of total probability that

$$\sum_{y\in\Omega} p(x,y) = p(x), \tag{1}$$

and

$$\sum_{x\in\Omega} p(x,y) = p(y). \tag{2}$$

Thus, we can then expand this definition as follows

$$\begin{aligned}
\mathbb{E}[X+Y] &= \sum_{x,y\in\Omega} (x+y)p(x,y) \\
&= \sum_{x,y\in\Omega} xp(x,y) + \sum_{x,y\in\Omega} yp(x,y) \\
&= \sum_{x\in\Omega} x \sum_{y\in\Omega} p(x,y) + \sum_{y\in\Omega} y \sum_{x\in\Omega} p(x,y) \\
&= \sum_{x\in\Omega} xp(x) + \sum_{y\in\Omega} yp(y) && \text{by (1) and (2)} \\
&= \mathbb{E}[X] + \mathbb{E}[Y] && \text{by definition} \\
\therefore \mathbb{E}[X+Y] &= \mathbb{E}[X] + \mathbb{E}[Y],
\end{aligned}$$

which proves the property we seek to prove.

(c) $\mathbb{E}[cX]$ where X is a discrete random variable and $c \in \mathbb{R}$ is a scalar that is not a random variable is given by

$$\mathbb{E}[cX] = c\mathbb{E}[X].$$

This one is more straightforward to prove and we again expand the definition as follows

$$\begin{aligned}
\mathbb{E}[cX] &= \sum_{x \in \Omega} cxp(x) \\
&= c \sum_{x \in \Omega} xp(x) \\
&= c\mathbb{E}[X] && \text{by definition} \\
\therefore \mathbb{E}[cX] &= c\mathbb{E}[X],
\end{aligned}$$

which proves the property we seek to prove.

(d) $\mathrm{Var}(X)$ where X is a discrete random variable is given by

$$\sum_{x \in \Omega} (x - \mathbb{E}[X])^2 p(x),$$

which can be also written as

$$\mathbb{E}[(X - \mathbb{E}[X])^2].$$

We also know that

$$\mathbb{E}[\mathbb{E}[X]] = \mathbb{E}[X], \tag{3}$$

and

$$\mathbb{E}[X\mathbb{E}[X]] = \mathbb{E}[X]\mathbb{E}[X] = \mathbb{E}[X]^2. \tag{4}$$

We then expand the definition as follows

$$\begin{aligned}
\mathrm{Var}(X) &= \sum_{x \in \Omega} (x - \mathbb{E}[X])^2 p(x) \\
&= \mathbb{E}[(X - \mathbb{E}[X])^2] \\
&= \mathbb{E}[(X^2 - 2X\mathbb{E}[X] + \mathbb{E}[X]^2)] \\
&= \mathbb{E}[X^2] - \mathbb{E}[2X\mathbb{E}[X]] + \mathbb{E}[\mathbb{E}[X]^2] && \text{by linearity} \\
&= \mathbb{E}[X^2] - 2\mathbb{E}[X]^2 + \mathbb{E}[\mathbb{E}[X]^2] && \text{by (3) and (4)} \\
&= \mathbb{E}[X^2] - 2\mathbb{E}[X]^2 + \mathbb{E}[X]^2 && \text{by (4)} \\
&= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \\
\therefore \mathrm{Var}(X) &= \mathbb{E}[X^2] - \mathbb{E}[X]^2,
\end{aligned}$$

which proves the property we seek to prove.

3. We define $M(x) = W^\top x$ where W is a weight matrix and x is an input vector. We can then use D_{tra} and the distance function D to define empirical cost function $\hat{R}$ as follows

$$\hat{R}(M, D_{\text{tra}}) = \frac{1}{N}\sum_{n=1}^{N} \|y_n^* - M(x_n)\|_2^2 = \frac{1}{N}\sum_{n=1}^{N} \left\|y_n^* - W^\top x_n\right\|_2^2 = \frac{1}{N}\sum_{n=1}^{N}(y_n^* - W^\top x_n)^2.$$

Thus, to minimize the emprical cost we can take the derivative of $\hat{R}$ with respect to W. First, we re-write $\hat{R}$ to make the differentiation easier

$$\frac{1}{N}\sum_{n=1}^{N}(y_n^* - W^\top x_n)^2 = \frac{1}{N}(Y - XW)^\top (Y - XW),$$

where $X \in \mathbb{R}^{d\times N}$ is the matrix of N input column vectors each with d "features" and $Y \in \mathbb{R}^{q\times N}$ is the matrix of N $q-$*dimensional* output column vectors. We now compute $\nabla_W \hat{R}$.

$$\begin{aligned}
\nabla_W \hat{R} &= \nabla_W(\frac{1}{N}(Y - XW)^\top (Y - XW)) && \\
&= \nabla_W(\frac{1}{N}(Y^\top - W^\top X^\top)(Y - XW)) && \text{evaluate transpose} \\
&= \nabla_W(\frac{1}{N}(Y^\top Y - Y^\top XW - W^\top X^\top Y + W^\top X^\top XW)) && \text{multiply terms} \\
&= \nabla_W(\frac{1}{N}(Y^\top Y - 2W^\top X^\top Y + W^\top X^\top XW)) && \text{collect like terms} \\
&= \frac{1}{N}(\nabla_W(Y^\top Y) - \nabla_W(2W^\top X^\top Y) + \nabla_W(W^\top X^\top XW)) && \text{by linearity} \\
&= \frac{1}{N}(-\nabla_W(2W^\top X^\top Y) + \nabla_W(W^\top X^\top XW)) && \text{evaluate gradient} \\
&= \frac{1}{N}(-\nabla_W(2W^\top X^\top Y) + \nabla_W((W^\top X^\top)(XW))) && \text{express as product} \\
&= \frac{1}{N}(-\nabla_W(2W^\top X^\top Y) + \nabla_W(W^\top X^\top)(XW) + (W^\top X^\top)\nabla_W(XW)) && \text{product rule} \\
&= \frac{1}{N}(-\nabla_W(2W^\top X^\top Y) + X^\top XW + W^\top X^\top X) && \text{evaluate gradient} \\
&= \frac{1}{N}(-\nabla_W(2W^\top X^\top Y) + 2X^\top XW) && \text{collect like terms} \\
&= \frac{1}{N}(-2X^\top Y + 2X^\top XW) && \text{evaluate gradient} \\
\therefore \nabla_W \hat{R} &= \frac{1}{N}(-2X^\top Y + 2X^\top XW) = 0 \implies X^\top XW = X^\top Y &&
\end{aligned}$$

Thus, we see that when $X^\top XW = X^\top Y$ then $\hat{R}$ is minimized. We can confirm this as the cost function is quadratic and convex due to the squaring of the ℓ_2 norm which implies that it has only a single extreme point and this point is a minimum. Assuming X has full rank, we can now use the Moore-Penrose pseudoinverse of X, which is given by $X^+ = (X^\top X)^{-1}X^\top$, and since $X^\top X$ is invertible (due to our assumption of full rank) we re-write our optimal solution as

$$W = (X^\top X)^{-1}X^\top Y,$$

and then use the definition of X^+ to get the optimal weight matrix

$$W = X^+ Y.$$

(Note: If X does not have full rank then X^+ can instead be computed using the SVD of X.)

4. (a) For notational convenience let $f = f(x)$ and $\hat{f} = \hat{f}(x;\Theta)$. We begin by considering the minimum L2 loss for a single example x where x is a random variable. We wish to find to find a $\hat{f}$ such that $\mathbb{E}_x$ is at a minimum which can be done by removing and adding f to the equation. We re-write the expectation and simplify the terms.

$$
\begin{aligned}
\mathbb{E}_x[(y-\hat{f})^2] &= \mathbb{E}_x[(y-f+f-\hat{f})^2] && \\
&= \mathbb{E}_x[((y-f)+(f-\hat{f}))^2] && \text{collect terms}\\
&= \mathbb{E}_x[(y-f)^2 - 2(y-f)(f-\hat{f}) + (f-\hat{f})^2] && \text{expand terms}\\
&= \mathbb{E}_x[(y-f)^2] - \mathbb{E}_x[2(y-f)(f-\hat{f})] + \mathbb{E}_x[(f-\hat{f})^2] && \text{by linearity}\\
&= \mathbb{E}_x[(y-f)^2] - \mathbb{E}_x[2(f+\epsilon-f)(f-\hat{f})] + \mathbb{E}_x[(f-\hat{f})^2] && \text{since } y = f+\epsilon\\
&= \mathbb{E}_x[(y-f)^2] - 2\mathbb{E}_x[(\epsilon)(f-\hat{f})] + \mathbb{E}_x[(f-\hat{f})^2] && \text{take out constant}\\
&= \mathbb{E}_x[(y-f)^2] - 2\mathbb{E}_x[\epsilon]\mathbb{E}_x[f-\hat{f}] + \mathbb{E}_x[(f-\hat{f})^2] && \text{since } \epsilon \text{ is a random}\\
& && \text{independent variable}\\
&= \mathbb{E}_x[(y-f)^2] + \mathbb{E}_x[(f-\hat{f})^2] && \text{since } \mathbb{E}_x[\epsilon] = 0\\
&= \mathbb{E}_x[(f+\epsilon-f)^2] + \mathbb{E}_x[(f-\hat{f})^2] && \text{since } y = f+\epsilon\\
&= \sigma^2 + \mathbb{E}_x[(f-\hat{f})^2] && \text{since } \mathbb{E}_x[\epsilon^2] = \sigma^2\\
& && \text{see derivation in (b)}
\end{aligned}
$$

$$\therefore \mathbb{E}_x[(y-\hat{f})^2] = \sigma^2 + \mathbb{E}_x[(f-\hat{f})^2]$$

Notice that the the value $f - \hat{f}$ is squared meaning whatever the value is, it will become positive and contribute positively to the constant σ^2. Thus, setting $\hat{f} = f$ means the second term will become zero and we will achieve a minimum L2 loss with the minimum loss being σ^2. We can then expand the single example case to a vector X and the distribution $Y = f(X) + \epsilon$ as the derivation of $\mathbb{E}_X$ will be identical to the derivation of $\mathbb{E}_x$ except this time we consider vectors. Since the L2 loss is minimized if the difference between each entry between X and $\hat{f}(X;\Theta)$ is minimized it is easy to see that choosing a $\hat{f}$ such that $\hat{f}(X;\Theta) = X$ will lead to a value of zero for the term $(f(X) - \hat{f}(X;\Theta))^2$ which again leads to a minimum L2 loss with the minimum loss being σ^2. Thus, we conclude that the minimum of L2 loss

$$\mathbb{E}_X[(Y - f(X;\Theta))^2]$$

is achieved for all x,

$$\hat{f}(x;\Theta) = f(x).$$

(b) For notational convenience let $f_0 = f(x_0)$ and $\hat{f}_0 = \hat{f}(x_0;\Theta)$. Then, note that $\mathrm{Var}[\hat{f}_0]$ is given by

$$\mathrm{Var}[\hat{f}_0] = \mathbb{E}[\hat{f}_0^2] - (\mathbb{E}[\hat{f}_0])^2, \tag{5}$$

and $\mathrm{Var}[y_0]$ is given by

$$\mathrm{Var}[y_0] = \mathbb{E}[y_0^2] - (\mathbb{E}[y_0])^2, \tag{6}$$

which we can then further simplify since we know that $\mathbb{E}[\epsilon] = 0$ (zero mean) and $\mathbb{E}(f_0) = f_0$ (since f_0 is deterministic.) This gives us

$$\mathbb{E}[y_0] = \mathbb{E}[f_0 + \epsilon] = \mathbb{E}[f_0] + \mathbb{E}[\epsilon] = f_0 + 0 = f_0,$$

which can then be used to simplify $\mathrm{Var}[y_0]$ by evaluating

$$\mathbb{E}[(y_0 - \mathbb{E}[y_0])^2] = \mathbb{E}[(f_0 + \epsilon - f_0)^2] = \mathbb{E}[\epsilon^2],$$

and then using the fact that $\mathrm{Var}[\epsilon] = \sigma^2$ (by the question construction) we can also deduce

$$\mathrm{Var}[\epsilon] = \mathbb{E}[\epsilon^2] - (\mathbb{E}[\epsilon])^2 = \mathbb{E}[\epsilon^2] = \sigma^2$$

which implies that $\mathrm{Var}[y_0] = \mathbb{E}[\epsilon^2] = \sigma^2$ which finally gives us

$$\mathrm{Var}[y_0] = \sigma^2. \tag{7}$$

Using our observations from above, we can then derive the bias variance decomposition as follows

$$\begin{aligned}
\mathbb{E}[(y_0 - \hat{f}_0)^2] &= \mathbb{E}[y_0^2 - 2y_0\hat{f}_0 + \hat{f}_0^2] \\
&= \mathbb{E}[y_0^2] - \mathbb{E}[2y_0\hat{f}_0] + \mathbb{E}[\hat{f}_0^2] && \text{by linearity} \\
&= \mathbb{E}[y_0^2] - \mathbb{E}[2y_0\hat{f}_0] + \mathrm{Var}[\hat{f}_0] + (\mathbb{E}[\hat{f}_0])^2 && \text{by (5)} \\
&= \mathrm{Var}[y_0] + (\mathbb{E}[y_0])^2 - \mathbb{E}[2y_0\hat{f}_0] + \mathrm{Var}[\hat{f}_0] + (\mathbb{E}[\hat{f}_0])^2 && \text{by (6)} \\
&= (\mathbb{E}[y_0])^2 - \mathbb{E}[2y_0\hat{f}_0] + (\mathbb{E}[\hat{f}_0])^2 + \mathrm{Var}[\hat{f}_0] + \mathrm{Var}[y_0] && \text{re-arranging terms} \\
&= f_0^2 - \mathbb{E}[2y_0\hat{f}_0] + (\mathbb{E}[\hat{f}_0])^2 + \mathrm{Var}[\hat{f}_0] + \mathrm{Var}[y_0] && \text{since } \mathbb{E}[y_0] = f_0 \\
&= f_0^2 - \mathbb{E}[2(f_0 + \epsilon)\hat{f}_0] + (\mathbb{E}[\hat{f}_0])^2 + \mathrm{Var}[\hat{f}_0] + \mathrm{Var}[y_0] && \text{since } y_0 = f_0 + \epsilon \\
&= f_0^2 - \mathbb{E}[2f_0\hat{f}_0 + 2\epsilon\hat{f}_0] + (\mathbb{E}[\hat{f}_0])^2 + \mathrm{Var}[\hat{f}_0] + \mathrm{Var}[y_0] && \text{expanding terms} \\
&= f_0^2 - \mathbb{E}[2f_0\hat{f}_0] + \mathbb{E}[2\epsilon\hat{f}_0] + (\mathbb{E}[\hat{f}_0])^2 + \mathrm{Var}[\hat{f}_0] + \mathrm{Var}[y_0] && \text{by linearity} \\
&= f_0^2 - \mathbb{E}[2f_0\hat{f}_0] + \mathbb{E}[\epsilon][2\hat{f}_0] + (\mathbb{E}[\hat{f}_0])^2 + \mathrm{Var}[\hat{f}_0] + \mathrm{Var}[y_0] && \text{since } \epsilon \text{ is a random} \\
& && \text{independent variable} \\
&= f_0^2 - \mathbb{E}[2f_0\hat{f}_0] + (\mathbb{E}[\hat{f}_0])^2 + \mathrm{Var}[\hat{f}_0] + \mathrm{Var}[y_0] && \text{since } \mathbb{E}[\epsilon] = 0 \\
&= f_0^2 - 2\mathbb{E}[f_0\hat{f}_0] + (\mathbb{E}[\hat{f}_0])^2 + \mathrm{Var}[\hat{f}_0] + \mathrm{Var}[y_0] && \text{take out constant} \\
&= f_0^2 - 2\mathbb{E}[f_0]\mathbb{E}[\hat{f}_0] + (\mathbb{E}[\hat{f}_0])^2 + \mathrm{Var}[\hat{f}_0] + \mathrm{Var}[y_0] && \text{since } f_0 \text{ and } \hat{f}_0 \text{ are} \\
& && \text{independent} \\
&= f_0^2 - 2f_0\mathbb{E}[\hat{f}_0] + (\mathbb{E}[\hat{f}_0])^2 + \mathrm{Var}[\hat{f}_0] + \mathrm{Var}[y_0] && \text{since } f_0 \text{ is determinsitic} \\
&= (f_0 - \mathbb{E}[\hat{f}_0])^2 + \mathrm{Var}[\hat{f}_0] + \mathrm{Var}[y_0] && \text{factor terms} \\
&= (\mathbb{E}[f_0] - \mathbb{E}[\hat{f}_0])^2 + \mathrm{Var}[\hat{f}_0] + \mathrm{Var}[y_0] && \text{since } f_0 = \mathbb{E}[f_0] \\
&= (\mathbb{E}[f_0 - \hat{f}_0])^2 + \mathrm{Var}[\hat{f}_0] + \mathrm{Var}[y_0] && \text{by linearity} \\
&= (\mathbb{E}[f_0 - \hat{f}_0])^2 + \mathrm{Var}[\hat{f}_0] + \sigma^2 && \text{by (7)} \\
\therefore \mathbb{E}[(y_0 - \hat{f}_0)^2] &= (\mathbb{E}[f_0 - \hat{f}_0])^2 + \mathrm{Var}[\hat{f}_0] + \sigma^2,
\end{aligned}$$

and then expanding on our notational convenience gives us the desired result

$$\mathbb{E}[(y_0 - \hat{f}(x_0; \Theta))^2] = (\mathbb{E}[f(x_0) - \hat{f}(x_0; \Theta)])^2 + \mathrm{Var}[\hat{f}(x_0; \Theta)] + \sigma^2$$

which shows the bias-variance decomposition.