1.

1.

The assumption $X \in \mathbb{R}^{N\times d}$ is a fixed data matrix means the rows $x_i^\top$ (the feature vectors) are not random variables in the probability model. All the randomness in the data-generating process comes only from the noise term $\vec{\epsilon}$. The features X are measured without error—i.e., they are "perfect" and are treated as constants. The only uncertainty comes from the outcome $\vec{y}$, via the noise term.
So the model is:

$$\vec{y} = X\vec{\theta} + \vec{\epsilon}, \quad \vec{\epsilon} \sim \mathcal{N}(0, \sigma^2 I_N)$$

Here, $\vec{y}$ is random, but X is treated as deterministic.
Ridge estimator:

$$\begin{aligned}
\hat{\vec{\theta}} &= (X^\top X + \lambda I_d)^{-1} X^\top \vec{y} \\
&= (X^\top X + \lambda I_d)^{-1} X^\top (X\vec{\theta} + \vec{\epsilon}) \\
&= (X^\top X + \lambda I_d)^{-1} X^\top X\vec{\theta} + (X^\top X + \lambda I_d)^{-1} X^\top \vec{\epsilon}
\end{aligned}$$

Because X is fixed, all expectations are conditional on X; the only randomness comes from $\vec{\epsilon}$. Thus $\mathbb{E}_{\mathcal{D}}[\hat{\vec{\theta}}] = \mathbb{E}[\hat{\vec{\theta}} \mid X]$.

$$\begin{aligned}
\mathbb{E}[\hat{\vec{\theta}} \mid X] &= (X^\top X + \lambda I_d)^{-1} X^\top X\vec{\theta} + (X^\top X + \lambda I_d)^{-1} X^\top \mathbb{E}[\vec{\epsilon} \mid X] \\
&= (X^\top X + \lambda I_d)^{-1} X^\top X\vec{\theta} + (X^\top X + \lambda I_d)^{-1} X^\top \cdot 0 \\
&\qquad \text{(Since } \vec{\epsilon} \sim \mathcal{N}(0, \sigma^2 I_N) \text{ and is independent of } X\text{, so } \mathbb{E}[\vec{\epsilon} \mid X] = \mathbb{E}[\vec{\epsilon}] = 0) \\
&= (X^\top X + \lambda I_d)^{-1} X^\top X\vec{\theta} \\
&= (I_d - \lambda (X^\top X + \lambda I_d)^{-1})\vec{\theta} \\
&\qquad \text{(Using } (A + \lambda I_d)^{-1} A = I_d - \lambda (A + \lambda I_d)^{-1} \text{ with } A = X^\top X)
\end{aligned}$$

Therefore,

$$\mathbb{E}_{\mathcal{D}}[\hat{\vec{\theta}}] = (I_d - \lambda (X^\top X + \lambda I_d)^{-1})\vec{\theta}$$

2.

No—ridge regression is biased (unless $\lambda = 0$). For $\hat{\vec{\theta}}$ to be unbiased, we would require

$$\mathbb{E}[\hat{\vec{\theta}}] = \vec{\theta}, \quad \text{for all } \vec{\theta}.$$

But instead we have

$$\mathbb{E}[\hat{\vec{\theta}}] = \big(I_d - \lambda (X^\top X + \lambda I_d)^{-1}\big)\vec{\theta}.$$

We pay a bias penalty (bias is non-zero) because of the ridge regularization (the ℓ_2 penalty), which shrinks coefficient estimates toward zero even when the linear model is correctly specified.

$$\mathrm{Bias}(\hat{\vec{\theta}}) = \mathbb{E}_{\mathcal{D}}[\hat{\vec{\theta}}] - \vec{\theta} = -\lambda (X^\top X + \lambda I_d)^{-1}\vec{\theta}.$$

This expression is the parameter bias: it measures the deviation of the expected ridge estimator from the true parameter vector $\vec{\theta}$.

Note: $\mathrm{Bias}(\hat{\vec{\theta}}) = 0 \iff \lambda = 0$ or $\vec{\theta} = 0$. The bias is non-zero for $\lambda > 0$ and $\vec{\theta} \neq 0$.

3.

Let $A := (X^\top X + \lambda I_d)^{-1} X^\top \in \mathbb{R}^{d\times N}$
Since X is fixed (deterministic), $\mathrm{Cov}(\hat{\vec{\theta}}) = \mathrm{Cov}(\hat{\vec{\theta}} \mid X)$.
Ridge estimator:

$$\begin{aligned}\hat{\vec{\theta}} &= (X^\top X + \lambda I_d)^{-1} X^\top \vec{y} \\ &= (X^\top X + \lambda I_d)^{-1} X^\top (X\vec{\theta} + \vec{\epsilon}) \\ &= (X^\top X + \lambda I_d)^{-1} X^\top X\vec{\theta} + (X^\top X + \lambda I_d)^{-1} X^\top \vec{\epsilon} \\ &= AX\vec{\theta} + A\vec{\epsilon}\end{aligned}$$

Expectation:

$$\begin{aligned}\mathbb{E}[\hat{\vec{\theta}}] &= (X^\top X + \lambda I_d)^{-1} X^\top X\vec{\theta} \\ &= AX\vec{\theta}\end{aligned}$$

Covariance derivation:

$$\begin{aligned}\mathrm{Cov}(\hat{\vec{\theta}}) &= \mathbb{E}\left[(\hat{\vec{\theta}} - \mathbb{E}[\hat{\vec{\theta}}])(\hat{\vec{\theta}} - \mathbb{E}[\hat{\vec{\theta}}])^\top\right] \\ &= \mathbb{E}\left[(AX\vec{\theta} + A\vec{\epsilon} - AX\vec{\theta})(AX\vec{\theta} + A\vec{\epsilon} - AX\vec{\theta})^\top\right] \\ &= \mathbb{E}\left[(A\vec{\epsilon})(A\vec{\epsilon})^\top\right] \\ &= \mathbb{E}\left[A\vec{\epsilon}\vec{\epsilon}^\top A^\top\right] \qquad \text{(Using } (AB)^\top = B^\top A^\top\text{)} \\ &= A\mathbb{E}[\vec{\epsilon}\vec{\epsilon}^\top]A^\top \\ &= A(\sigma^2 I_N)A^\top \qquad \text{(As } \mathbb{E}[\vec{\epsilon}\vec{\epsilon}^\top] = \sigma^2 I_N\text{, since } \vec{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_N)\text{)} \\ &= \sigma^2 AA^\top\end{aligned}$$

Substituting back $A = (X^\top X + \lambda I_d)^{-1} X^\top$:

$$\begin{aligned}\mathrm{Cov}(\hat{\vec{\theta}}) &= \sigma^2 AA^\top \\ &= \sigma^2 (X^\top X + \lambda I_d)^{-1} X^\top \left[(X^\top X + \lambda I_d)^{-1} X^\top\right]^\top \\ &= \sigma^2 (X^\top X + \lambda I_d)^{-1} X^\top X (X^\top X + \lambda I_d)^{-T} \\ &= \sigma^2 (X^\top X + \lambda I_d)^{-1} X^\top X (X^\top X + \lambda I_d)^{-1}\end{aligned}$$

Note: $X^\top X + \lambda I_d$ is symmetric positive definite (for $\lambda > 0$), so $(X^\top X + \lambda I_d)^{-T} = (X^\top X + \lambda I_d)^{-1}$.

4.

We consider the ridge regression setting with training data:

$$\vec{y} = X\vec{\theta} + \vec{\varepsilon}, \qquad \vec{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_N)$$

where $X \in \mathbb{R}^{N\times d}$ is fixed (deterministic) and $\vec{\theta} \in \mathbb{R}^d$ is the true parameter vector.

For a fixed test point $\vec{x}^{(0)} \in \mathbb{R}^d$, the true response is:

$$y^{(0)} = f(\vec{x}^{(0)}) + \varepsilon^{(0)} = \vec{x}^{(0)\top}\vec{\theta} + \varepsilon^{(0)}, \qquad \varepsilon^{(0)} \sim \mathcal{N}(0, \sigma^2)$$

where $\varepsilon^{(0)}$ is independent of the training data.

The ridge estimator from training data $\mathcal{D} = (X, \vec{y})$ is:

$$\hat{\vec{\theta}} = (X^\top X + \lambda I_d)^{-1} X^\top \vec{y}$$

The prediction at the test point $\vec{x}^{(0)}$ is:

$$\hat{y}^{(0)} = \vec{x}^{(0)\top}\hat{\vec{\theta}}$$

The EPE at test point $\vec{x}^{(0)}$ is:

$$\begin{aligned}
\mathrm{EPE}(\vec{x}^{(0)}) &= \mathbb{E}_{\mathcal{D},y^{(0)}}\left[\left(y^{(0)} - \hat{y}^{(0)}\right)^2\right] \\
&= \mathbb{E}_{\mathcal{D},\varepsilon^{(0)}}\left[\left(f(\vec{x}^{(0)}) + \varepsilon^{(0)} - \hat{y}^{(0)}\right)^2\right] \\
&= \mathbb{E}_{\mathcal{D},\varepsilon^{(0)}}\left[\left(f(\vec{x}^{(0)}) - \hat{y}^{(0)} + \varepsilon^{(0)}\right)^2\right] \\
&= \mathbb{E}_{\mathcal{D},\varepsilon^{(0)}}\left[\left(f(\vec{x}^{(0)}) - \hat{y}^{(0)}\right)^2 + 2\varepsilon^{(0)}\left(f(\vec{x}^{(0)}) - \hat{y}^{(0)}\right) + \left(\varepsilon^{(0)}\right)^2\right] \\
&= \mathbb{E}_{\mathcal{D},\varepsilon^{(0)}}\left[\left(f(\vec{x}^{(0)}) - \hat{y}^{(0)}\right)^2\right] + \mathbb{E}_{\mathcal{D}}\left[f(\vec{x}^{(0)}) - \hat{y}^{(0)}\right] \cdot \mathbb{E}_{\varepsilon^{(0)}}\left[2\varepsilon^{(0)}\right] + \mathbb{E}_{\varepsilon^{(0)}}\left[\left(\varepsilon^{(0)}\right)^2\right] \\
&= \mathbb{E}_{\mathcal{D}}\left[\left(f(\vec{x}^{(0)}) - \hat{y}^{(0)}\right)^2\right] + \sigma^2 \qquad (\text{As } \varepsilon^{(0)} \sim \mathcal{N}(0,\sigma^2), \text{ so } \mathbb{E}[\varepsilon^{(0)}] = 0 \text{ and } \mathbb{E}[(\varepsilon^{(0)})^2] = \sigma^2) \\
&= \left(f(\vec{x}^{(0)}) - \mathbb{E}_{\mathcal{D}}[\hat{y}^{(0)}]\right)^2 + \mathrm{Var}_{\mathcal{D}}(\hat{y}^{(0)}) + \sigma^2 \quad (\text{Using } \mathbb{E}[(a-Z)^2] = (a - \mathbb{E}[Z])^2 + \mathrm{Var}(Z))
\end{aligned}$$

Therefore:

$$\mathrm{EPE}(\vec{x}^{(0)}) = \sigma^2 + \left(f(\vec{x}^{(0)}) - \mathbb{E}_{\mathcal{D}}[\hat{y}^{(0)}]\right)^2 + \mathrm{Var}_{\mathcal{D}}(\hat{y}^{(0)})$$

For ridge with fixed $X \in \mathbb{R}^{N\times d}$:

Let $S = (X^\top X + \lambda I_d)^{-1} \in \mathbb{R}^{d\times d}$.

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}[\hat{\vec{\theta}}] &= (I_d - \lambda(X^\top X + \lambda I_d)^{-1})\vec{\theta} \\
&= (I_d - \lambda S)\vec{\theta}
\end{aligned}$$

$$\begin{aligned}
\mathrm{Cov}(\hat{\vec{\theta}}) &= \sigma^2(X^\top X + \lambda I_d)^{-1}X^\top X(X^\top X + \lambda I_d)^{-1} \\
&= \sigma^2 S X^\top X S
\end{aligned}$$

This gives us:

$$\begin{aligned}
\mathbb{E}[\hat{\vec{\theta}}] &= SX^\top X\vec{\theta} = (I_d - \lambda S)\vec{\theta} \\
\mathrm{Cov}(\hat{\vec{\theta}}) &= \sigma^2 S X^\top X S
\end{aligned}$$

Irreducible Error: σ^2 (noise variance)
Bias: Using prediction space. At test point $\vec{x}^{(0)}$, the prediction is $\hat{y}^{(0)} = \vec{x}^{(0)\top}\hat{\vec{\theta}}$.

$$\begin{aligned}
\mathbb{E}[\hat{y}^{(0)}] &= \vec{x}^{(0)\top}\mathbb{E}[\hat{\vec{\theta}}] = \vec{x}^{(0)\top}(I_d - \lambda S)\vec{\theta} \\
\mathrm{Bias} &= f(\vec{x}^{(0)}) - \mathbb{E}[\hat{y}^{(0)}] \\
&= \vec{x}^{(0)\top}\vec{\theta} - \vec{x}^{(0)\top}(I_d - \lambda S)\vec{\theta} \\
&= \lambda\vec{x}^{(0)\top}S\vec{\theta} \\
\mathrm{Bias}^2 &= \lambda^2\vec{\theta}^\top S\vec{x}^{(0)}\vec{x}^{(0)\top}S\vec{\theta}
\end{aligned}$$

Variance: Using $\mathrm{Var}(a^\top Z) = a^\top \mathrm{Cov}(Z)a$ with $a = \vec{x}^{(0)}$ and $Z = \hat{\vec{\theta}}$:

$$\begin{aligned}
\mathrm{Var}(\hat{y}^{(0)}) &= \mathrm{Var}(\vec{x}^{(0)\top}\hat{\vec{\theta}}) = \vec{x}^{(0)\top}\,\mathrm{Cov}(\hat{\vec{\theta}})\vec{x}^{(0)} \\
&= \sigma^2\vec{x}^{(0)\top}SX^\top XS\vec{x}^{(0)}
\end{aligned}$$

Final decomposition:

$$\text{EPE}(\vec{x}^{(0)}) = \sigma^2 + \lambda^2 \vec{\theta}^\top S \vec{x}^{(0)} \vec{x}^{(0)\top} S \vec{\theta} + \sigma^2 \vec{x}^{(0)\top} S X^\top X S \vec{x}^{(0)}$$

Substituting back $S = (X^\top X + \lambda I_d)^{-1}$:

$$\text{EPE}(\vec{x}^{(0)}) = \sigma^2 + \lambda^2 \vec{\theta}^\top (X^\top X + \lambda I_d)^{-1} \vec{x}^{(0)} \vec{x}^{(0)\top} (X^\top X + \lambda I_d)^{-1} \vec{\theta} + \sigma^2 \vec{x}^{(0)\top} (X^\top X + \lambda I_d)^{-1} X^\top X (X^\top X + \lambda I_d)^{-1} \vec{x}^{(0)}$$

2.

1.

The assumption $X \in \mathbb{R}^{N \times d}$ is a fixed data matrix with rank d means the rows $x_i^\top$ (the feature vectors) are not random variables in the probability model. All the randomness in the data-generating process comes only from the noise term $\vec{\epsilon}$. The features X are measured without error—i.e., they are "perfect" and are treated as constants. The only uncertainty comes from the outcome $\vec{y}$, via the noise term.
Assume the data was drawn from

$$y = h(\vec{x}) + \epsilon = \vec{\theta}^\top \vec{x} + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2).$$

An estimate of $h(\vec{x})$ is given by

$$\hat{h}(\vec{x}) = \vec{\beta}^\top X \vec{x},$$

where the parameter vector $\vec{\beta} \in \mathbb{R}^N$.
The cost function is defined as:

$$J(\vec{\beta}) := \sum_{j=1}^{N} \left(\hat{h}(\vec{x}) - y^{(i)} \right)^2 + \lambda \|X^\top \vec{\beta}\|^2,$$

that is

$$J(\vec{\beta}) := (XX^\top \vec{\beta} - \vec{y})^\top (XX^\top \vec{\beta} - \vec{y}) + \vec{\beta}^\top X \lambda X^\top \vec{\beta}$$

Let $K := XX^\top \in \mathbb{R}^{N \times N}$

$$\begin{aligned} J(\vec{\beta}) &= (XX^\top \vec{\beta} - \vec{y})^\top (XX^\top \vec{\beta} - \vec{y}) + \vec{\beta}^\top X \lambda X^\top \vec{\beta} \\ &= (K\vec{\beta} - \vec{y})^\top (K\vec{\beta} - \vec{y}) + \lambda \vec{\beta}^\top K \vec{\beta} \\ &= (\vec{\beta}^\top K^\top - \vec{y}^\top)(K\vec{\beta} - \vec{y}) + \lambda \vec{\beta}^\top K \vec{\beta} \\ &= \vec{\beta}^\top K^\top K \vec{\beta} - \vec{\beta}^\top K^\top \vec{y} - \vec{y}^\top K \vec{\beta} + \vec{y}^\top \vec{y} + \lambda \vec{\beta}^\top K \vec{\beta} \end{aligned}$$

Using $K = K^\top$ (since $K = XX^\top$ is symmetric):

$$\begin{aligned} J(\vec{\beta}) &= \vec{\beta}^\top K K \vec{\beta} - \vec{\beta}^\top K \vec{y} - \vec{y}^\top K \vec{\beta} + \vec{y}^\top \vec{y} + \lambda \vec{\beta}^\top K \vec{\beta} \\ &= \vec{\beta}^\top K^2 \vec{\beta} - 2\vec{y}^\top K \vec{\beta} + \vec{y}^\top \vec{y} + \lambda \vec{\beta}^\top K \vec{\beta} \end{aligned}$$

Taking the gradient with respect to $\vec{\beta}$:

$$\begin{aligned} \nabla_\beta J(\vec{\beta}) &= \nabla_\beta \left[\vec{\beta}^\top K^2 \vec{\beta} - 2\vec{y}^\top K \vec{\beta} + \vec{y}^\top \vec{y} + \lambda \vec{\beta}^\top K \vec{\beta} \right] \\ &= \nabla_\beta \left[\vec{\beta}^\top K^2 \vec{\beta} \right] - \nabla_\beta \left[2\vec{y}^\top K \vec{\beta} \right] + \nabla_\beta \left[\vec{y}^\top \vec{y} \right] + \nabla_\beta \left[\lambda \vec{\beta}^\top K \vec{\beta} \right] \\ &= (K^2 + (K^2)^\top)\vec{\beta} - 2K^\top \vec{y} + 0 + \lambda (K + K^\top)\vec{\beta} \quad (\text{As } \nabla_\beta(\vec{\beta}^\top A \vec{\beta}) = (A + A^\top)\vec{\beta} \text{ and } \nabla_\beta(\vec{c}^\top \vec{\beta}) = \vec{c}) \\ &= 2K^2 \vec{\beta} - 2K\vec{y} + 2\lambda K \vec{\beta} \quad (\text{As } K = K^\top \Rightarrow K^2 = (K^2)^\top) \\ &= 2(K^2 \vec{\beta} + \lambda K \vec{\beta} - K\vec{y}) \\ &= 2K(K + \lambda I)\vec{\beta} - 2K\vec{y} \end{aligned}$$

Therefore:

$$\nabla_\beta J(\vec{\beta}) = 2K(K+\lambda I)\vec{\beta} - 2K\vec{y}$$

$$\begin{aligned} \nabla_\beta J(\vec{\beta}) &= 0 \\ \Rightarrow 2K(K+\lambda I)\vec{\beta} - 2K\vec{y} &= 0 \\ \Rightarrow K(K+\lambda I)\vec{\beta} &= K\vec{y} \end{aligned}$$

For any vector $\vec{v} \in \mathbb{R}^N$, we have:

$$\begin{aligned} \vec{v}^\top K \vec{v} &= \vec{v}^\top (XX^\top)\vec{v} \\ &= (\vec{v}^\top X)(X^\top \vec{v}) \quad (\text{As } (AB)C = A(BC)) \\ &= (X^\top \vec{v})^\top (X^\top \vec{v}) \quad (\text{As } \vec{v}^\top X = (X^\top \vec{v})^\top) \\ &= \|X^\top \vec{v}\|^2 \geq 0 \quad (\text{As } \|\vec{u}\|^2 = \vec{u}^\top \vec{u} \geq 0) \end{aligned}$$

Since $\vec{v}^\top K \vec{v} \geq 0$ for all $\vec{v} \in \mathbb{R}^N$, the matrix $K = XX^\top$ is positive semi-definite.
For $\lambda > 0$, the matrix $(K + \lambda I)$ is positive definite because:

$$\vec{v}^\top (K+\lambda I)\vec{v} = \vec{v}^\top K \vec{v} + \lambda \|\vec{v}\|^2 > 0 \quad \text{for all } \vec{v} \neq \vec{0}$$

Therefore $(K + \lambda I)$ is invertible.
The solution $\hat{\vec{\beta}} = (K+\lambda I)^{-1}\vec{y}$ satisfies the optimality condition:

$$\begin{aligned} K(K+\lambda I)\hat{\vec{\beta}} &= K(K+\lambda I)(K+\lambda I)^{-1}\vec{y} \\ &= K\vec{y} \end{aligned}$$

Therefore, the optimal solution that minimizes the cost function is:

$$\hat{\vec{\beta}} = (K+\lambda I)^{-1}\vec{y} = (XX^\top + \lambda I)^{-1}\vec{y}$$

2.

We want the degrees of freedom:

$$\mathrm{df}(\hat{y}) = \frac{1}{\sigma^2}\mathrm{tr}\big(\mathrm{Cov}(\hat{y}, y)\big).$$

From the optimization problem, the solution is

$$\hat{\beta} = (K+\lambda I)^{-1} y, \qquad K = XX^\top.$$

Therefore, the fitted values are

$$\hat{y} = K\hat{\beta} = K(K+\lambda I)^{-1} y.$$

This can be written as

$$\hat{y} = Ay, \qquad A := K(K+\lambda I)^{-1}.$$

As $y = X\theta + \varepsilon$ with $X\theta$ fixed (non-random), all randomness in y comes from ε. Hence

$$\mathrm{Cov}(y, y) = \mathrm{Cov}(\varepsilon, \varepsilon) = \sigma^2 I.$$

Using $\hat{y} = Ay$,

$$\mathrm{Cov}(\hat{y}, y) = \mathrm{Cov}(Ay, y) = A\,\mathrm{Cov}(y, y) = A\sigma^2 I.$$

Therefore,

$$\mathrm{df}(\hat{y}) = \frac{1}{\sigma^2}\mathrm{tr}(A\sigma^2 I) = \mathrm{tr}(A).$$

Substituting back $A = K(K+\lambda I)^{-1}$:

$$\mathrm{df}(\hat{y}) = \mathrm{tr}\left(K(K+\lambda I)^{-1}\right)$$