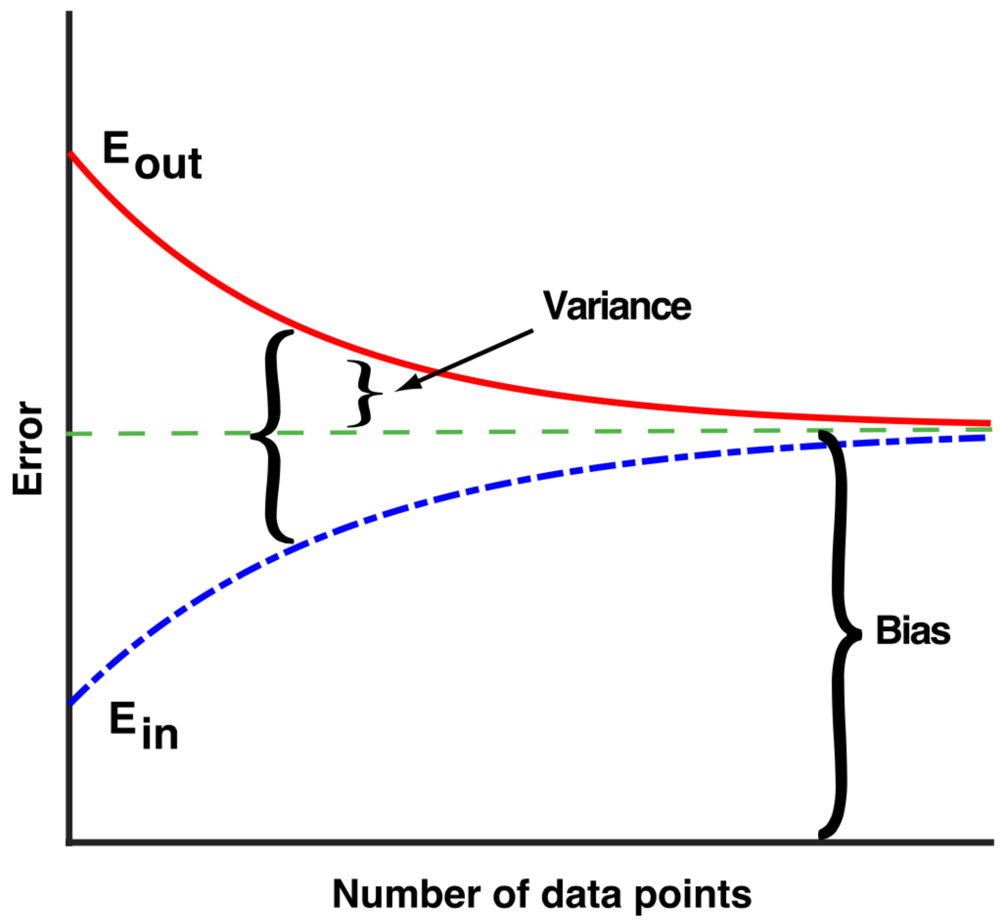
Machine Learning for Physical Scientists

Lecture 3

Regularization: a simple way to reduce generalization error

Recap: Linear Regression and Generalization Error

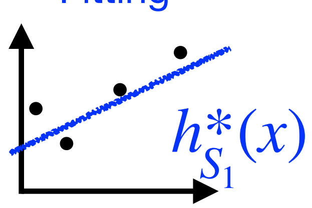


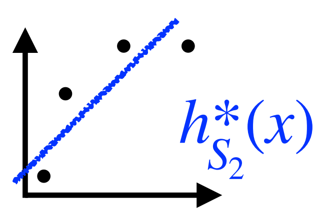
$$\bar{E}_{in} \equiv \mathbb{E}_D \left[E_{in}(\boldsymbol{w}^*_D) \right] = \sigma^2 \left(1 - \frac{d}{m} \right)$$ (in-sample error)

$$\bar{E}_{out} \equiv \mathbb{E}_D \left[E_{out}(\boldsymbol{w}^*_D) \right] = \sigma^2 \left(1 + \frac{d}{m} \right)$$ (out-of-sample error)

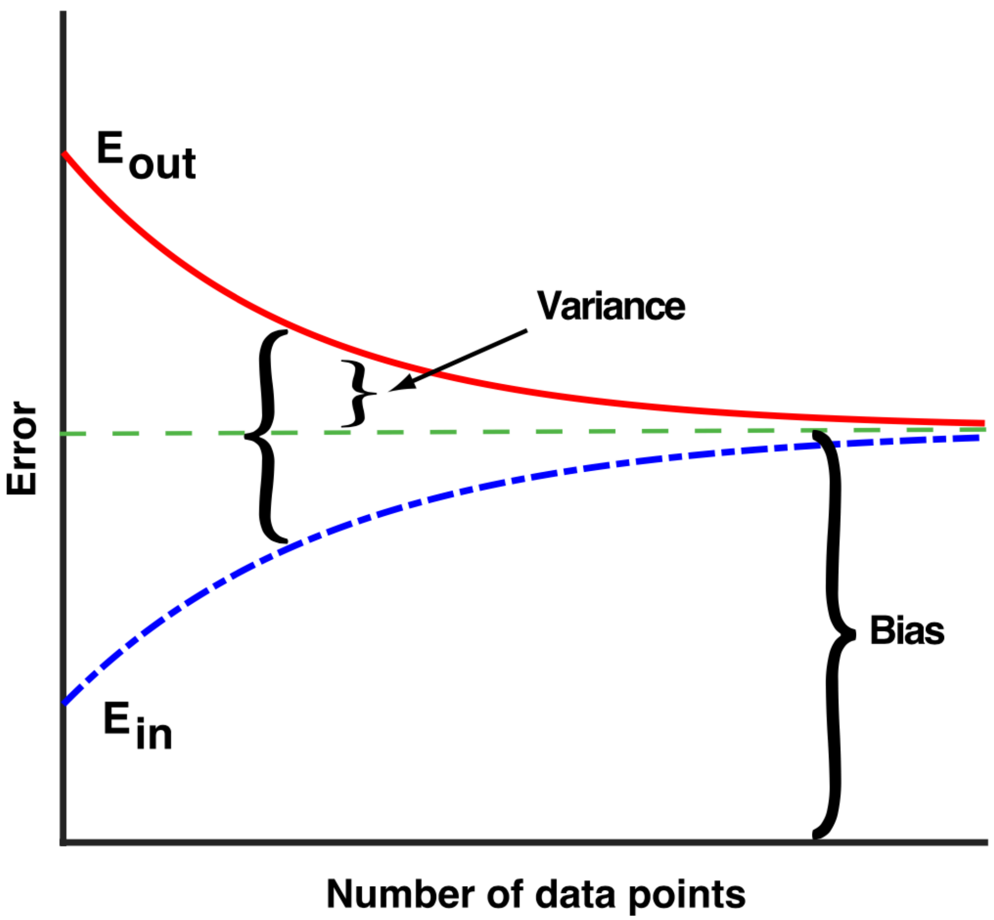
$$|\bar{E}_{out} - \bar{E}_{in}| = 2\sigma^2 \left(\frac{d}{m} \right)$$ generalization error

Fitting





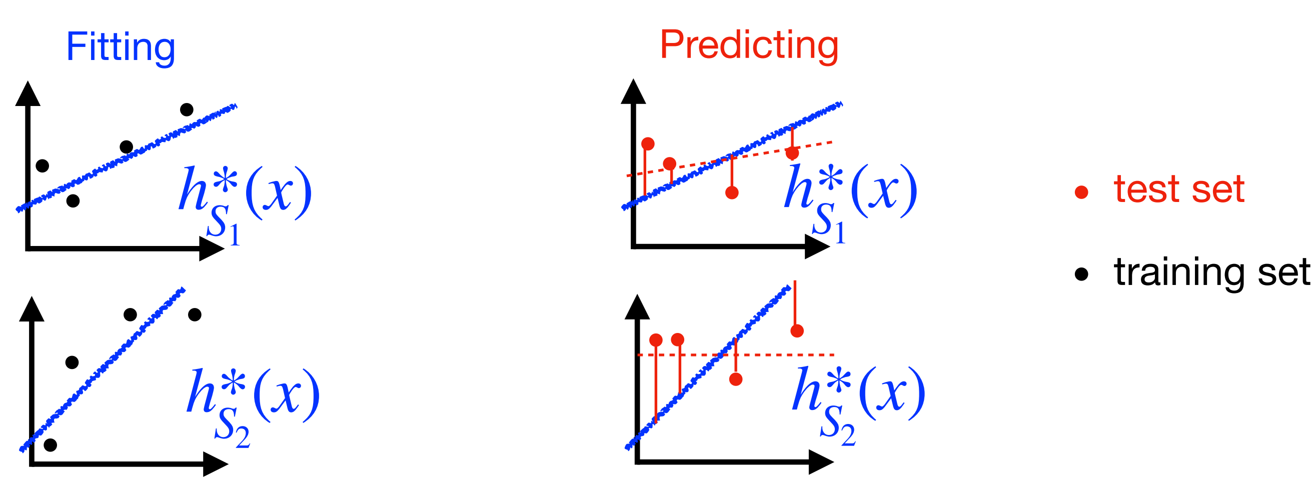
Recap: Linear Regression and Generalization Error



$$\bar{E}_{in} \equiv \mathbb{E}_D\left[E_{in}(\boldsymbol{w}^*_D)\right] = \sigma^2\left(1 - \frac{d}{m}\right)$$ (in-sample error)

$$\bar{E}_{out} \equiv \mathbb{E}_D\left[E_{out}(\boldsymbol{w}^*_D)\right] = \sigma^2\left(1 + \frac{d}{m}\right)$$ (out-of-sample error)

$$|\,\bar{E}_{out} - \bar{E}_{in}\,| = 2\sigma^2\left(\frac{d}{m}\right)$$ generalization error



You'll likely fit the noise rather than signal in a small sample size limit!

How shall we regularise the model to not be too sensitive to new data in the limit of small sample size?

1. Regularization

2. Validation

ordinary least square

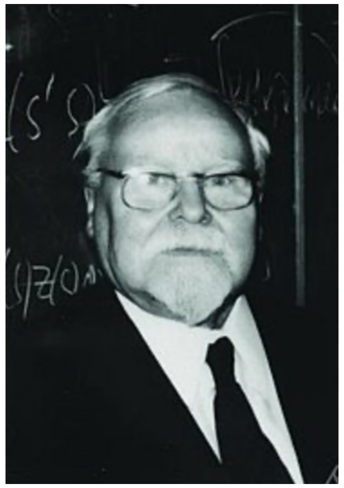
$$L_S(\boldsymbol{w}_{ls}^*) = \frac{1}{m} \min_{\boldsymbol{w} \in \mathbb{R}^d} \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2$$

ordinary least square

$$L_S(\boldsymbol{w}^*_{ls}) = \frac{1}{m} \min_{\boldsymbol{w}\in\mathbb{R}^d} \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2$$

regularized least square

$$L_S(\boldsymbol{w}^*_{ridge}) \equiv \min_{\boldsymbol{w}\in\mathbb{R}^d} \left[\frac{1}{m}\sum_{i=1}^{m}\left(y_i - \boldsymbol{w}^T\boldsymbol{x}^{(i)}\right)^2 + \lambda \boldsymbol{w}^T\boldsymbol{w}\right], \quad \lambda \geq 0$$



Tikhonov ‘62

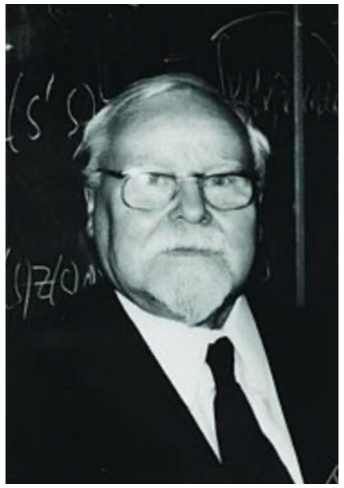
ordinary least square

$$L_S(\boldsymbol{w}^*_{ls}) = \frac{1}{m} \min_{\boldsymbol{w}\in\mathbb{R}^d} \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2$$

regularized least square

$$L_S(\boldsymbol{w}^*_{ridge}) \equiv \min_{\boldsymbol{w}\in\mathbb{R}^d} \left[\frac{1}{m}\sum_{i=1}^{m}\left(y_i - \boldsymbol{w}^T\boldsymbol{x}^{(i)}\right)^2 + \lambda \boldsymbol{w}^T\boldsymbol{w}\right], \quad \lambda \geq 0$$

$$= \min_{\boldsymbol{w}\in\mathbb{R}^d} \Bigg[\underbrace{\frac{1}{m}\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2}_{\text{standard empirical risk}} + \underbrace{\lambda\|\boldsymbol{w}\|_2^2}_{\text{regularizer (risk penalty)}}\Bigg], \qquad \lambda \geq 0$$



Tikhonov ‘62

Soft-constraint, rather than setting some directions to be 0.

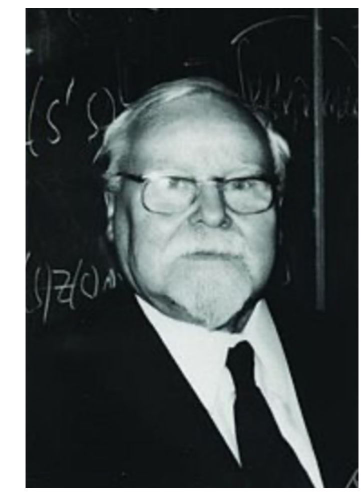
ordinary least square

$$L_S(\boldsymbol{w}^*_{ls}) = \frac{1}{m} \min_{\boldsymbol{w}\in\mathbb{R}^d} \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2$$

regularized least square

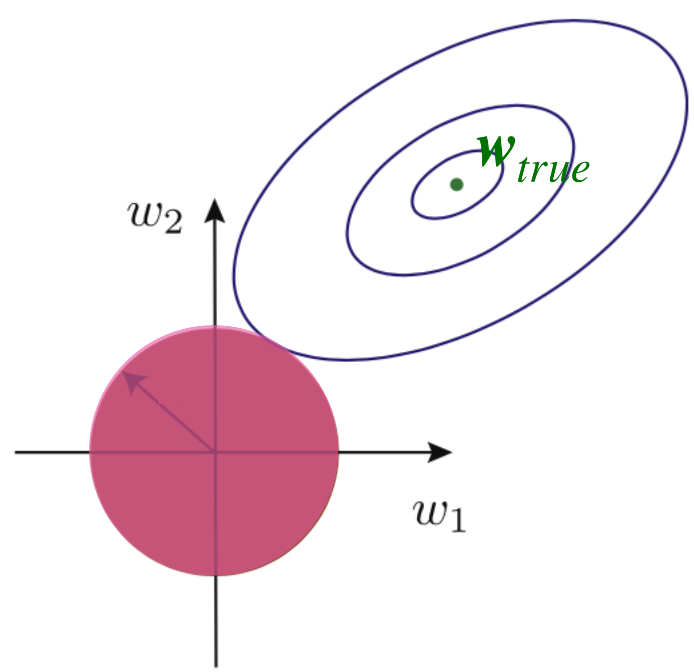
$$L_S(\boldsymbol{w}^*_{ridge}) \equiv \min_{\boldsymbol{w}\in\mathbb{R}^d} \left[\frac{1}{m}\sum_{i=1}^{m}\left(y_i - \boldsymbol{w}^T\boldsymbol{x}^{(i)}\right)^2 + \lambda \boldsymbol{w}^T\boldsymbol{w}\right], \quad \lambda \geq 0$$

$$= \min_{\boldsymbol{w}\in\mathbb{R}^d} \left[\underbrace{\frac{1}{m}\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2}_{\text{standard empirical risk}} + \underbrace{\lambda\|\boldsymbol{w}\|_2^2}_{\text{regularizer (risk penalty)}}\right], \qquad \lambda \geq 0$$



Tikhonov '62

Soft-constraint, rather than setting some directions to be 0.



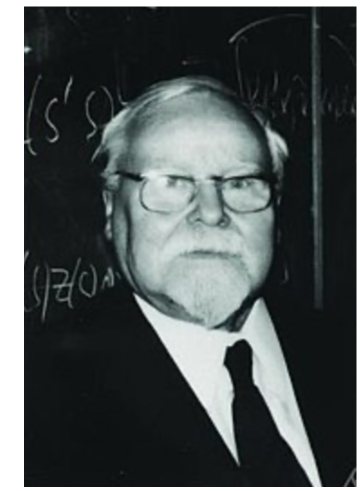
ordinary least square

$$L_S(\boldsymbol{w}^*_{ls}) = \frac{1}{m} \min_{\boldsymbol{w} \in \mathbb{R}^d} \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2$$

regularized least square

$$L_S(\boldsymbol{w}^*_{ridge}) \equiv \min_{\boldsymbol{w} \in \mathbb{R}^d} \left[\frac{1}{m} \sum_{i=1}^{m} \left(y_i - \boldsymbol{w}^T \boldsymbol{x}^{(i)} \right)^2 + \lambda \boldsymbol{w}^T \boldsymbol{w} \right], \quad \lambda \geq 0$$

$$= \min_{\boldsymbol{w} \in \mathbb{R}^d} \Bigg[\underbrace{\frac{1}{m} \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2}_{\text{standard empirical risk}} + \underbrace{\lambda \|\boldsymbol{w}\|_2^2}_{\text{regularizer (risk penalty)}} \Bigg], \qquad \lambda \geq 0$$



Tikhonov ‘62

Let’s see how the critical point (which is also the minimizer since this is a convex optimization problem) depends on the data. As usual, we’ll take the gradient of the loss function above and set to zero:

$$\left(\boldsymbol{X}^T \boldsymbol{X} + \lambda \boldsymbol{I}_m \right) \boldsymbol{w}^*_{ridge} = \boldsymbol{X}^T \boldsymbol{y}$$

Assuming invertibility, we get

$$\boldsymbol{w}^*_{ridge} = \left(\boldsymbol{X}^T \boldsymbol{X} + \lambda \boldsymbol{I}_m \right)^{-1} \boldsymbol{X}^T \boldsymbol{y}$$

We'll perform Singular Value Decomposition (SVD) to see how each component of $\boldsymbol{y}^*_{ridge}$ is related to $\boldsymbol{y}^*_{ls}$.

Recall that any matrix $\boldsymbol{X} \in \mathbb{R}^{m \times d}$ can be decomposed into the product of orthogonal matrices $\boldsymbol{U} \in \mathbb{R}^{m \times d}$, $\boldsymbol{V} \in \mathbb{R}^{d \times d}$, and the diagonal matrix $\Sigma = diag(\sigma_1, \sigma_2, \ldots, \sigma_d)$, whose diagonals are the singular values of $\boldsymbol{X}$ such that $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_d \geq 0$, as

$$\boldsymbol{X} = \boldsymbol{U}\Sigma\boldsymbol{V}^T .$$

We'll perform Singular Value Decomposition (SVD) to see how each component of $\boldsymbol{y}^*_{ridge}$ is related to $\boldsymbol{y}^*_{ls}$.

Recall that any matrix $\boldsymbol{X} \in \mathbb{R}^{m\times d}$ can be decomposed into the product of orthogonal matrices $\boldsymbol{U} \in \mathbb{R}^{m\times d}$, $\boldsymbol{V} \in \mathbb{R}^{d\times d}$, and the diagonal matrix $\Sigma = diag(\sigma_1, \sigma_2, \ldots, \sigma_d)$, whose diagonals are the singular values of $\boldsymbol{X}$ such that $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_d \geq 0$, as

$$\boldsymbol{X} = \boldsymbol{U}\Sigma\boldsymbol{V}^T .$$

Performing the decomposition, one gets

$$\begin{aligned}\boldsymbol{w}^*_{ridge} &= \left(\boldsymbol{X}^T\boldsymbol{X}+\lambda\boldsymbol{I}_m\right)^{-1}\boldsymbol{X}^T\boldsymbol{y} \\ &= \boldsymbol{V}\left(\Sigma^2+\lambda\boldsymbol{I}_m\right)^{-1}\Sigma\boldsymbol{U}^T\boldsymbol{y}\end{aligned}$$

We'll perform Singular Value Decomposition (SVD) to see how each component of $\boldsymbol{y}^*_{ridge}$ is related to $\boldsymbol{y}^*_{ls}$.

Recall that any matrix $\boldsymbol{X} \in \mathbb{R}^{m \times d}$ can be decomposed into the product of orthogonal matrices $\boldsymbol{U} \in \mathbb{R}^{m \times d}$, $\boldsymbol{V} \in \mathbb{R}^{d \times d}$, and the diagonal matrix $\Sigma = diag(\sigma_1, \sigma_2, \ldots, \sigma_d)$, whose diagonals are the singular values of $\boldsymbol{X}$ such that $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_d \geq 0$, as

$$\boldsymbol{X} = \boldsymbol{U}\Sigma\boldsymbol{V}^T.$$

Performing the decomposition, one gets

$$\begin{aligned}\boldsymbol{w}^*_{ridge} &= \left(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}_m\right)^{-1}\boldsymbol{X}^T\boldsymbol{y} \\ &= \boldsymbol{V}\left(\Sigma^2 + \lambda\boldsymbol{I}_m\right)^{-1}\Sigma\boldsymbol{U}^T\boldsymbol{y}\end{aligned}$$

This gives us

$$\begin{aligned}\boldsymbol{y}^*_{ridge} &= \boldsymbol{X}\boldsymbol{w}^*_{ridge} \\ &= \boldsymbol{U}\Sigma\left(\Sigma^2 + \lambda\boldsymbol{I}_m\right)^{-1}\Sigma\boldsymbol{U}^{\mathrm{T}}\boldsymbol{y}\end{aligned}$$

We'll perform Singular Value Decomposition (SVD) to see how each component of $\boldsymbol{y}^*_{ridge}$ is related to $\boldsymbol{y}^*_{ls}$.

Recall that any matrix $\boldsymbol{X} \in \mathbb{R}^{m\times d}$ can be decomposed into the product of orthogonal matrices $\boldsymbol{U} \in \mathbb{R}^{m\times d}$, $\boldsymbol{V} \in \mathbb{R}^{d\times d}$, and the diagonal matrix $\Sigma = diag(\sigma_1, \sigma_2, \ldots, \sigma_d)$, whose diagonals are the singular values of $\boldsymbol{X}$ such that $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_d \geq 0$, as

$$\boldsymbol{X} = \boldsymbol{U}\Sigma\boldsymbol{V}^T .$$

Performing the decomposition, one gets

$$\begin{aligned} \boldsymbol{w}^*_{ridge} &= \left(\boldsymbol{X}^T\boldsymbol{X}+\lambda\boldsymbol{I}_m\right)^{-1}\boldsymbol{X}^T\boldsymbol{y} \\ &= \boldsymbol{V}\left(\Sigma^2+\lambda\boldsymbol{I}_m\right)^{-1}\Sigma\boldsymbol{U}^T\boldsymbol{y} \end{aligned}$$

This gives us

$$\begin{aligned} \boldsymbol{y}^*_{ridge} &= \boldsymbol{X}\boldsymbol{w}^*_{ridge} \\ &= \boldsymbol{U}\Sigma\left(\Sigma^2+\lambda\boldsymbol{I}_m\right)^{-1}\Sigma\boldsymbol{U}^{\mathrm{T}}\boldsymbol{y} \\ &= \sum_{i=1}^{d}\boldsymbol{U}_{:,i}\frac{\sigma_i^2}{\sigma_i^2+\lambda}\boldsymbol{U}_{:i}^{\mathrm{T}}\boldsymbol{y} \end{aligned}$$

Compare to $\lambda = 0$ which is the result of standard least square, we can see that the size of the regularized ridge regression prediction is constrained by λ.

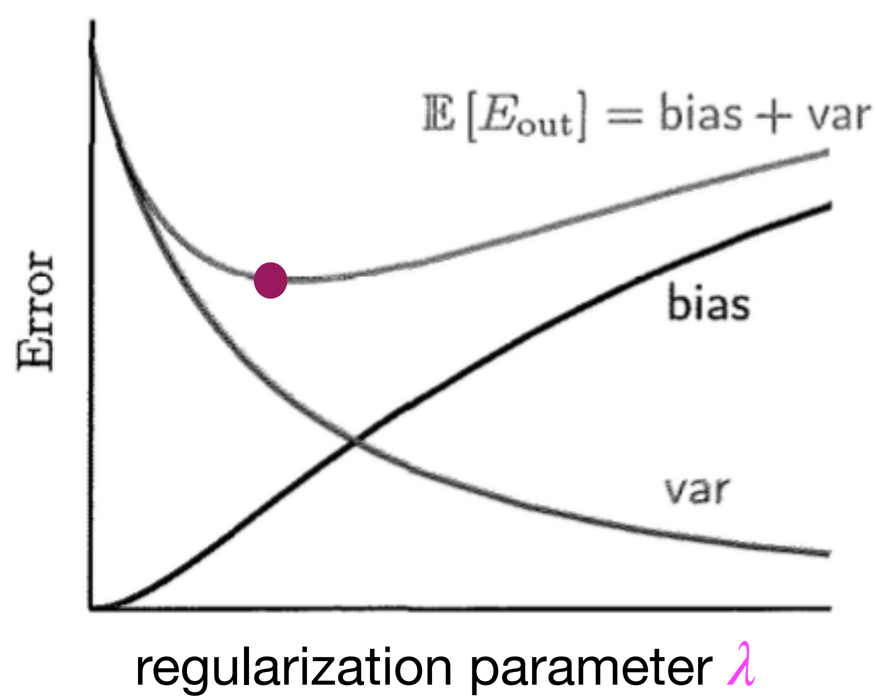
Can one be more quantitative about the generalization error?

In fact, for ridge regression of linear least square, you will derive the following fact (homework1) :

In the asymptotic limit in which $m \gg 1$,

$$bias(\lambda) \approx \frac{\lambda^2}{(\lambda+m)^2} \| \boldsymbol{w}_{true} \|_2^2,$$

$$var(\lambda) \approx \frac{\sigma^2}{(1+\lambda)^2} \left(\frac{d}{m}\right)$$

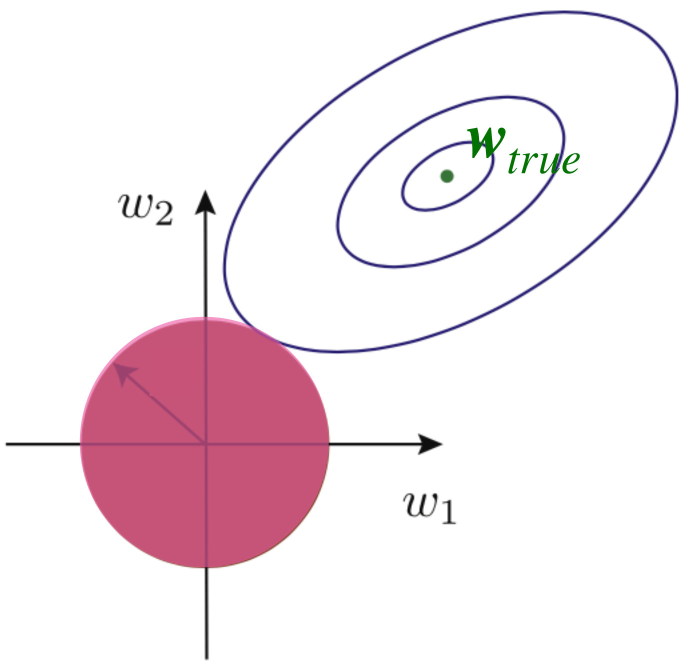


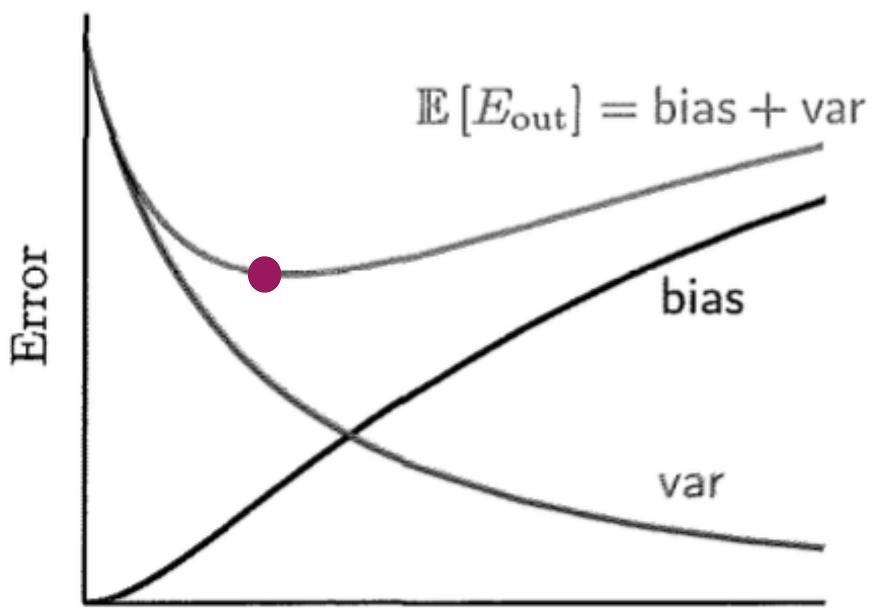
In fact, for ridge regression of linear least square, you will derive the following fact (homework1) :

In the asymptotic limit in which $m \gg 1$,

$$bias(\lambda) \approx \frac{\lambda^2}{(\lambda+m)^2} \| \boldsymbol{w}_{true} \|_2^2,$$

$$var(\lambda) \approx \frac{\sigma^2}{(1+\lambda)^2} \left(\frac{d}{m}\right)$$





regularization parameter λ