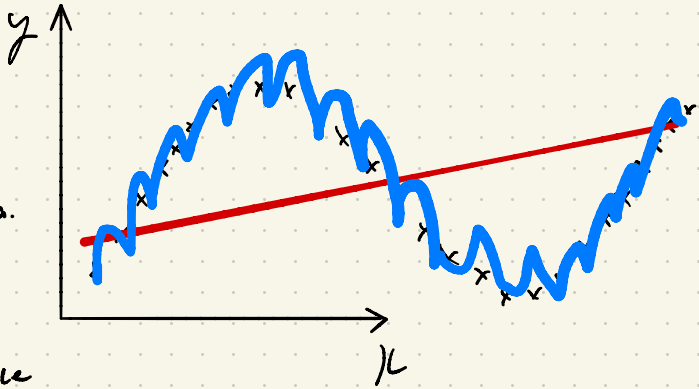
Bias Variance decomposition

- Non-linear Regression

$x \in \mathbb{R}$

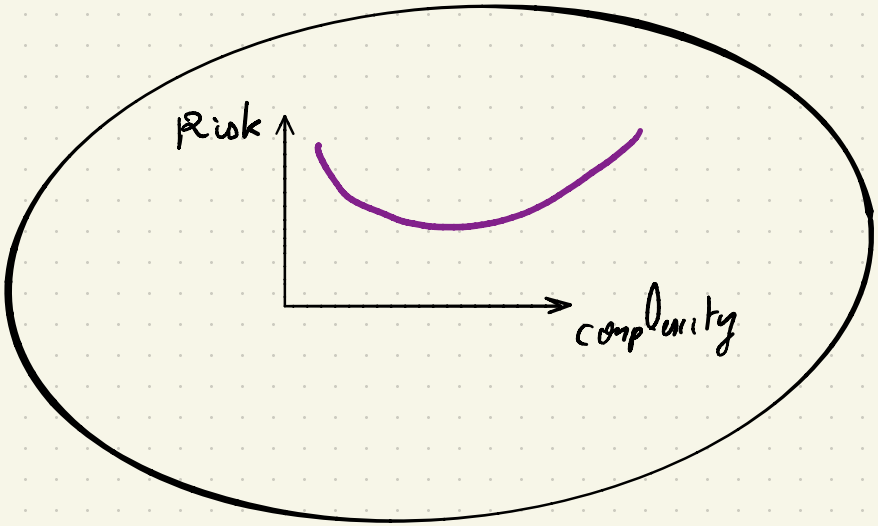
$(x, x^2, x^3, \dots, x^d)$

↑ complexity

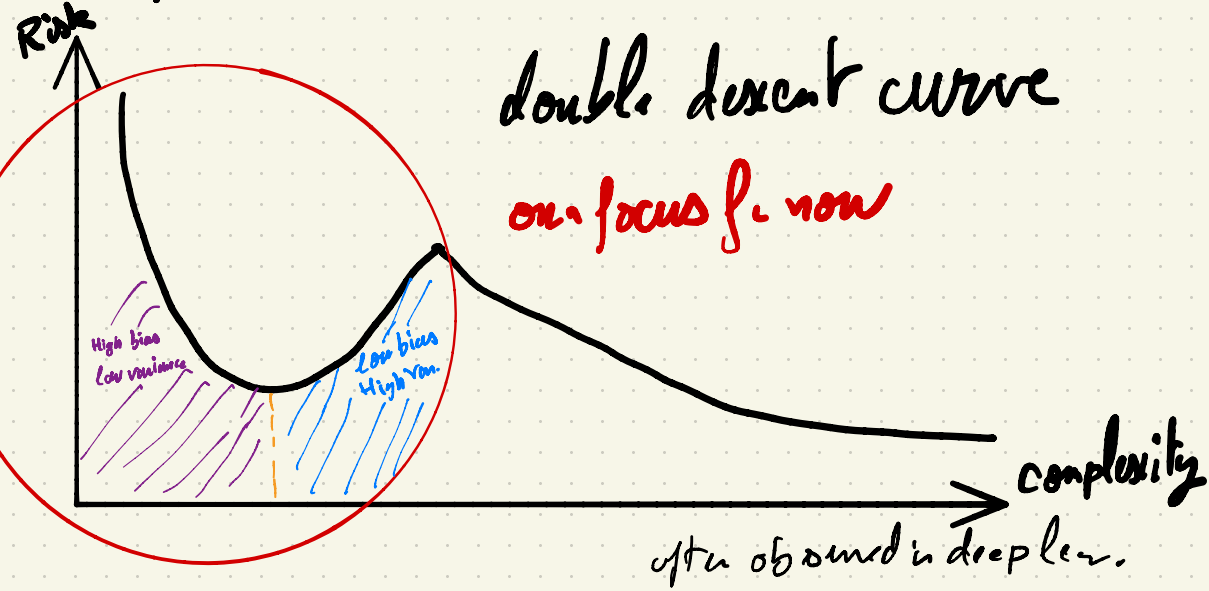


complexity of the class:

- low complexity → • large bias
  • small varia.
- High complexity → • low bias
  • High variance



Important note



double descent curve

our focus for now

often observed in deep learn.

Data model:

$$y = f(x) + \varepsilon$$

$f(x)$: true model

ε: noise

noise: $\varepsilon \sim D_\varepsilon$ iid, $\perp\!\!\!\perp D_x$

$\mathbb{E}[\varepsilon] = 0$

- In general, the model is not realizable.
  i.e. f is not in our model class

- $\mathbb{E}_{S_{Train}}\left[\left(y - f_{S_{Train}}(x)\right)^2\right]$

$S_{Train} \rightarrow$ Algo. to get $f_{S_{Train}} = A(S_{Train})$

x_0 fixed $\quad (y - \hat{f}_{S_{Train}}(x_0))^2 = (f(x_0) + \varepsilon - \hat{f}_{S_{Train}}(x_0))^2$

$$\mathbb{E}_{\substack{\varepsilon \sim D_\varepsilon \\ S_{Tr} \sim D}}\left[(y - \hat{f}_{S_{Train}}(x_0))^2\right] = \mathbb{E}\left[(f(x_0) + \varepsilon - \hat{f}_{S_{Train}}(x_0))^2\right]$$

$$= \mathbb{E}_{S_{Train} \sim D}\left[\left(f(x_0) - \hat{f}_{S_{Train}}(x_0)\right)^2\right] + \mathbb{E}_{\varepsilon \sim D_\varepsilon}\left[\varepsilon^2\right] + \text{Cross Term}$$

Since $\mathbb{E}[\varepsilon] = 0$; $\mathbb{E}[\varepsilon^2] = Var(\varepsilon)$

Cross Term: $\mathbb{E}\left[2\varepsilon\left(f(x_0) - \hat{f}_{S_{Train}}(x_0)\right)\right] = 2\,\mathbb{E}[\varepsilon]\,\mathbb{E}\left[f(x_0) - \hat{f}_{S_{Train}}(x_0)\right]$ ($D_\varepsilon \perp\!\!\!\perp D$, $\mathbb{E}[\varepsilon] = 0$) $= 0$

$$\mathbb{E}_{\substack{\varepsilon \sim D_\varepsilon \\ S_{Tr} \sim D}}\left[\left(y - \hat{f}_{S_{Tr}}(x_0)\right)^2\right] = Var(\varepsilon) + \mathbb{E}_{S_{Tr}}\left[\left(f(x_0) - \hat{f}_{S_{Tr}}(x_0)\right)^2\right]$$

$$\mathbb{E}_{S_{Tr}}\left[\left(f(x_0) - \hat{f}_{S_{Tr}}(x_0)\right)^2\right] = \mathbb{E}_{S_{Tr} \sim D}\left[\left(f(x_0) - \mathbb{E}_{S'_{Tr} \sim D}\left[\hat{f}_{S'_{Train}}(x_0)\right] + \mathbb{E}_{S'_{Tr} \sim D_{Tr}}\left[\hat{f}_{S'_{Tr}}(x_0)\right] - \hat{f}_{S_{Tr}}(x_0)\right)^2\right]$$

$$= \left(f(x_0) - \mathbb{E}_{S'_{Tr}}\left[\hat{f}_{S'_{Train}}(x_0)\right]\right)^2 + \mathbb{E}_{S_{Train}}\left[\left(\hat{f}_{S_{Train}}(x_0) - \mathbb{E}_{S'_{Tr}}\left(\hat{f}_{S'_{Tr}}(x_0)\right)\right)^2\right] + \text{Cross Term}$$

C.T $= \mathbb{E}_{S_{Train}}\left[\left(f(x_0) - \mathbb{E}_{S'_{Train}}\left[\hat{f}_{S'_{Train}}(x_0)\right]\right) \cdot \left(\mathbb{E}_{S'_{Train}}\left[\hat{f}_{S'_{Train}}(x_0)\right] - \hat{f}_{S_{Train}}(x_0)\right)\right]$

$= \left(f(x_0) - \mathbb{E}_{S'}\left(\hat{f}_{S'}(x_0)\right)\right) \cdot \left(\mathbb{E}_{S'}\left[\hat{f}_{S'}(x_0)\right] - \mathbb{E}_{S}\left[\hat{f}_{S}(x_0)\right]\right) = 0$

finally

$$\mathbb{E}_{\substack{\varepsilon \\ S_{Train}}}\left[\left(f(x_0)+\varepsilon-f_{S_{Train}}(x_0)\right)^2\right] \qquad (y-f(x_0))^2$$

$$= Var(\varepsilon) + \left(f(x_0) - \mathbb{E}_{S_{Train}}\left[f_{S_{Train}}(x_0)\right]\right)^2$$

bias

$$+ \mathbb{E}_{S_{Train}}\left[\left(f_{S_{Train}}(x_0) - \mathbb{E}_{S_{Train}}\left[f_{S_{Train}}(x_0)\right]\right)^2\right]$$

Variance

$\Rightarrow$ Decomposition in 3 positive terms:

- Noise Term, noise data
- Bias Term
- Variance Term

Bias Term:

$f(x_0)$ = actual function

$\mathbb{E}_{S_{Train}}\left[f_{S_{Train}}(x_0)\right]$: expectation of prediction (average predict.)

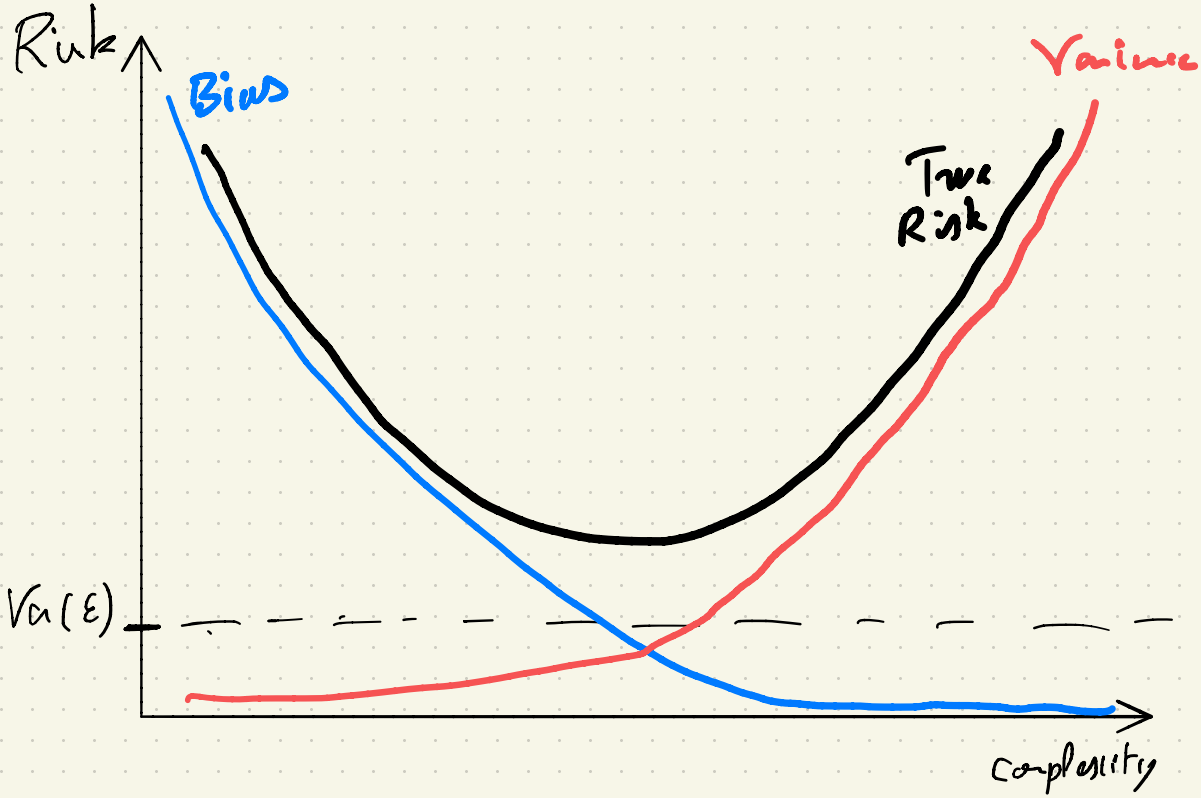
$\rightarrow$ if complexity small $\rightarrow$ high bias
High $\rightarrow$ small bias

Variance Term :

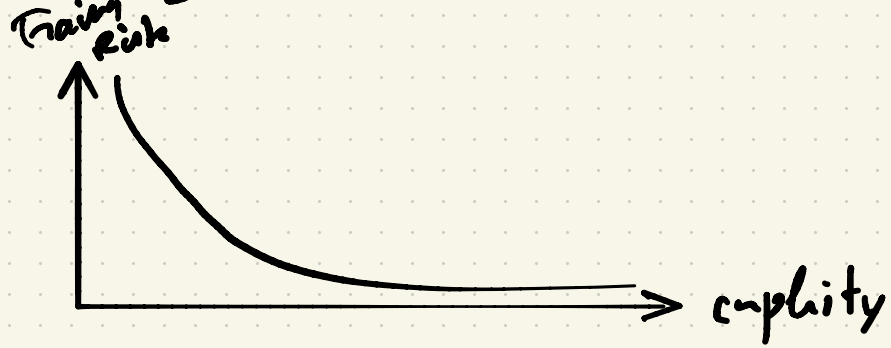
consistency of prediction

Complexity ↗ Variance ↗

In summary :



Training risk :



- For neural networks: Train with a var. alg. (SGD)