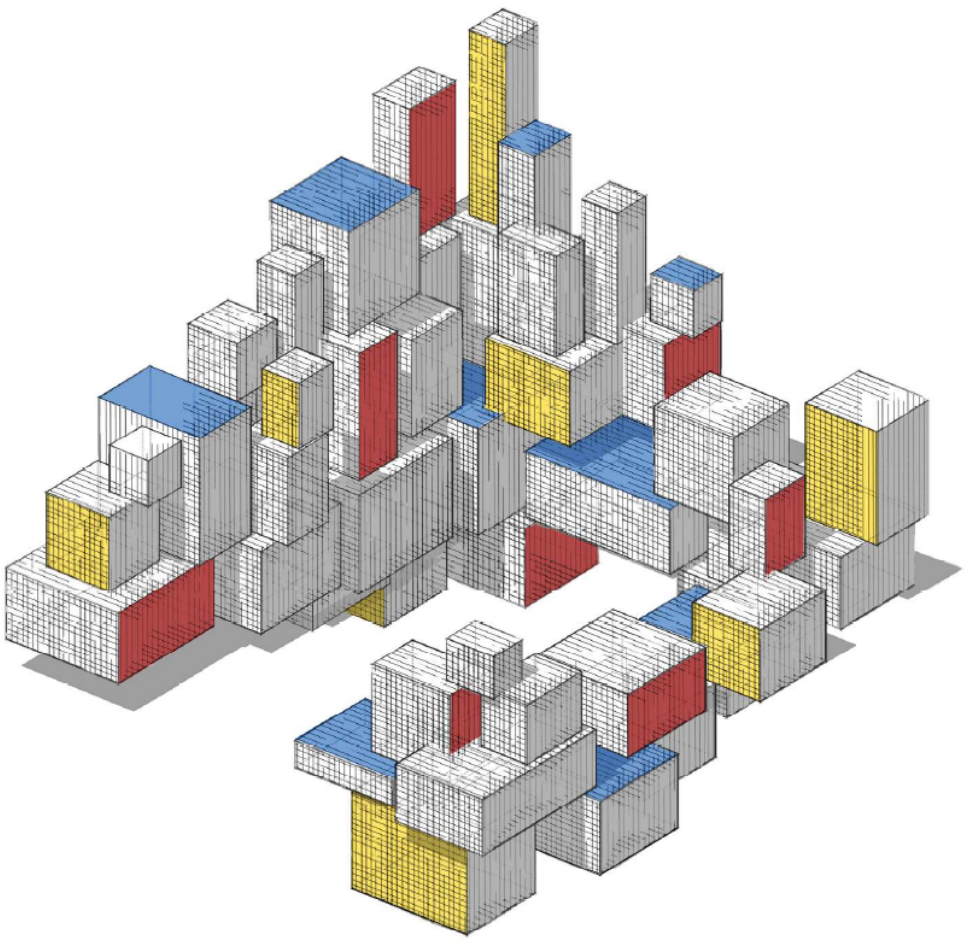
Tradeoffs in Statistical Learning



Purpose

In this lecture we discuss:

- Decomposition of the generalization risk into:
  - irreducible risk
  - approximation error
  - statistical error
- Approximation-estimation tradeoff
- Bias-variance tradeoff

Tradeoffs in Statistical Learning

We wish to make the (expected) generalization risk as small as possible, while using as few computational resources as possible.

First, a suitable class $\mathcal{G}$ of prediction functions has to be chosen, driven by

- the complexity of the class (e.g., is it rich enough to adequately approximate, or even contain, the optimal prediction function g^*?),
- the ease of training the learner via optimizing the training loss,
- how accurately the training loss estimates the risk within class $\mathcal{G}$,
- the feature types (categorical, continuous, etc.).

Tradeoffs in Statistical Learning

The choice of a suitable function class $\mathcal{G}$ involves a tradeoff between conflicting factors.

For example, a learner from a simple class $\mathcal{G}$ can be trained very quickly, but may not approximate g^* very well, whereas a learner from a rich class $\mathcal{G}$ that contains g^* may require a lot of computing resources to train.

To better understand the relation between model complexity, computational simplicity, and estimation accuracy, it is useful to decompose the generalization risk into several parts.

Decomposing the Generalization Risk

Recall that for a training set τ, prediction function class $\mathcal{G}$, and learner $g_\tau^{\mathcal{G}}$, the generalization risk is the expected loss $\ell(g_\tau^{\mathcal{G}}) = \mathbb{E}\text{Loss}(Y, g_\tau^{\mathcal{G}}(\boldsymbol{X}))$.

We can decompose the generalization risk into the following three components:

$$\ell(g_\tau^{\mathcal{G}}) = \underbrace{\ell^*}_{\text{irreducible risk}} + \underbrace{\ell(g^{\mathcal{G}}) - \ell^*}_{\text{approximation error}} + \underbrace{\ell(g_\tau^{\mathcal{G}}) - \ell(g^{\mathcal{G}})}_{\text{statistical error}}, \qquad (1)$$

where $\ell^* := \ell(g^*)$ is the irreducible risk and $g^{\mathcal{G}} := \text{argmin}_{g \in \mathcal{G}} \ell(g)$ is the best learner within class $\mathcal{G}$. No learner can predict a new response with a smaller risk than ℓ^*.

Approximation error

The approximation error $\ell(g^{\mathcal{G}}) - \ell^*$ measures the difference between the irreducible risk and the best possible risk that can be obtained within function class $\mathcal{G}$.

Determining a suitable class $\mathcal{G}$ and minimizing $\ell(g)$ over this class is purely a problem of numerical and functional analysis, as the training data τ are not present.

For a fixed $\mathcal{G}$ that does not contain the optimal g^*, the approximation error cannot be made arbitrarily small and may be the dominant component in the generalization risk. The only way to reduce the approximation error then is by expanding the class $\mathcal{G}$ to include a larger set of possible functions.

Statistical Error

The statistical (estimation) error $\ell(g_\tau^{\mathcal{G}}) - \ell(g^{\mathcal{G}})$ depends on the training set τ and, in particular, on how well the learner $g_\tau^{\mathcal{G}}$ estimates the best possible prediction function, $g^{\mathcal{G}}$, within class $\mathcal{G}$.

For any sensible estimator this error should decay to zero (in probability or expectation) as the training size tends to infinity.

The approximation–estimation tradeoff pits two competing demands:

- The class $\mathcal{G}$ has to be simple enough so that the statistical error is not too large.
- The class $\mathcal{G}$ has to be rich enough to ensure a small approximation error.

Thus, there is a tradeoff between the approximation and estimation errors.

Squared-error Loss Decomposition

For the special case of the squared-error loss, the generalization risk is equal to

$$\ell(g_\tau^{\mathcal{G}}) = \mathbb{E}(Y - g_\tau^{\mathcal{G}}(\boldsymbol{X}))^2.$$

Recall that in this case the optimal prediction function is given by $g^*(\boldsymbol{x}) = \mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}]$. The decomposition (1) can now be interpreted as follows.

1. The irreducible error is $\ell^* = \mathbb{E}(Y - g^*(\boldsymbol{X}))^2$.
2. The approximation error is $\mathbb{E}(g^{\mathcal{G}}(\boldsymbol{X}) - g^*(\boldsymbol{X}))^2$.
3. For the statistical error $\ell(g_\tau^{\mathcal{G}}) - \ell(g^{\mathcal{G}})$ there is no direct interpretation as an expected squared error, *unless* $\mathcal{G}$ is the class of linear functions; that is, $g(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\beta}$ for some vector $\boldsymbol{\beta}$.
   In this case, we can write the statistical error as

$$\ell(g_\tau^{\mathcal{G}}) - \ell(g^{\mathcal{G}}) = \mathbb{E}(g_\tau^{\mathcal{G}}(\boldsymbol{X}) - g^{\mathcal{G}}(\boldsymbol{X}))^2.$$

Squared-error Loss for Linear Prediction Functions

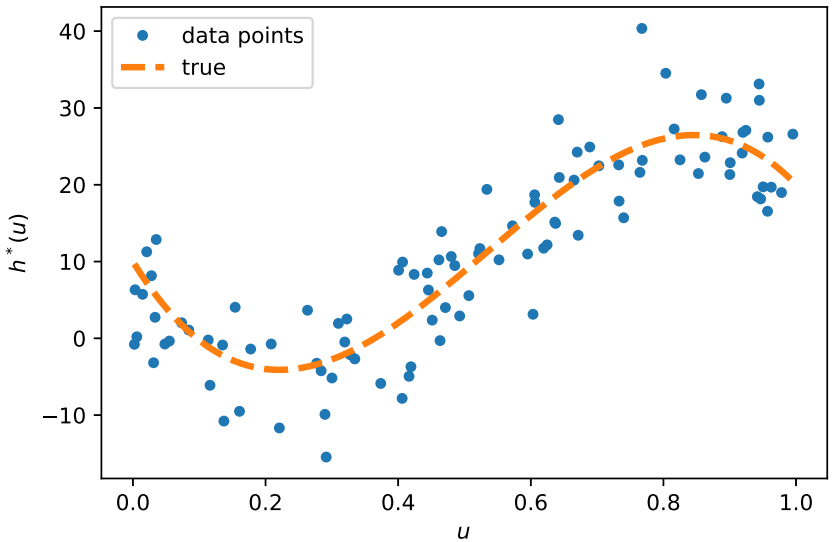
Thus, when using a squared-error loss, the generalization risk for a linear class $\mathcal{G}$ can be decomposed as:

$$\ell(g_\tau^{\mathcal{G}}) = \mathbb{E}(g_\tau^{\mathcal{G}}(\boldsymbol{X}) - Y)^2 = \ell^* + \underbrace{\mathbb{E}(g^{\mathcal{G}}(\boldsymbol{X}) - g^*(\boldsymbol{X}))^2}_{\text{approximation error}} + \underbrace{\mathbb{E}(g_\tau^{\mathcal{G}}(\boldsymbol{X}) - g^{\mathcal{G}}(\boldsymbol{X}))^2}_{\text{statistical error}}.$$

Note that in this decomposition the statistical error is the only term that depends on the training set.

Example: Polynomial Regression (cont.)

Consider the polynomial regression example: $\{U_i\} \sim_{\text{iid}} \mathcal{U}(0,1)$ and $(Y_i \mid U_u = u_i) \sim \mathcal{N}(10 - 140u_i + 400u_i^2 - 250u_i^3,\ 25)$.



We use feature vectors of the form $\boldsymbol{x} = [1, u, u^2, \ldots, u^{p-1}]^\top$. Let $\mathcal{G} = \mathcal{G}_p$ be the class of linear functions of such vectors and let $g^*(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\beta}^*$. Conditional on $\boldsymbol{X} = \boldsymbol{x}$, we have:

$$Y = g^*(\boldsymbol{x}) + \varepsilon(\boldsymbol{x}), \quad \varepsilon(\boldsymbol{x}) \sim \mathcal{N}(0, \ell^*),$$

where $\ell^* = \mathbb{E}(Y - g^*(\boldsymbol{X}))^2 = 25$ is the irreducible error.

Approximation Error

We wish to understand how the approximation error behaves as we change the complexity parameter p.

Any function $g \in \mathcal{G}_p$ can be written as

$$g(\boldsymbol{x}) = [1, u, \ldots, u^{p-1}]\,\boldsymbol{\beta},$$

and so $g(\boldsymbol{X})$ is distributed as $[1, U, \ldots, U^{p-1}]\boldsymbol{\beta}$, where $U \sim \mathcal{U}(0, 1)$.

Similarly, $g^*(\boldsymbol{X})$ is distributed as $[1, U, U^2, U^3]\boldsymbol{\beta}^*$.

It follows that an expression for the approximation error is:

$$\int_0^1 \left([1, u, \ldots, u^{p-1}]\,\boldsymbol{\beta} - [1, u, u^2, u^3]\,\boldsymbol{\beta}^* \right)^2 \mathrm{d}u.$$

Approximation Error

To minimize this error, we set the gradient with respect to $\boldsymbol{\beta}$ to zero and obtain the p linear equations:

$$\begin{aligned}
\int_0^1 \left([1, u, \ldots, u^{p-1}]\,\boldsymbol{\beta} - [1, u, u^2, u^3]\,\boldsymbol{\beta}^*\right)\,\mathrm{d}u &= 0,\\
\int_0^1 \left([1, u, \ldots, u^{p-1}]\,\boldsymbol{\beta} - [1, u, u^2, u^3\,]\boldsymbol{\beta}^*\right)\,u\,\mathrm{d}u &= 0,\\
&\vdots\\
\int_0^1 \left([1, u, \ldots, u^{p-1}]\,\boldsymbol{\beta} - [1, u, u^2, u^3]\,\boldsymbol{\beta}^*\right)\,u^{p-1}\mathrm{d}u &= 0.
\end{aligned}$$

This can be written as the matrix equation

$$\mathbf{H}_p\boldsymbol{\beta} = \widetilde{\mathbf{H}}\boldsymbol{\beta}^*,$$

where $\mathbf{H}_p = \int_0^1 [1, u, \ldots, u^{p-1}]^\top [1, u, \ldots, u^{p-1}]\,\mathrm{d}u$ is a $p \times p$ Hilbert matrix and $\widetilde{\mathbf{H}}$ is the $p \times 4$ upper-left block of $\mathbf{H}_{\widetilde{p}}$, with $\widetilde{p} = \max\{p, 4\}$.

Approximation Error

The solution, $\boldsymbol{\beta}_p$, is:

$$\boldsymbol{\beta}_p = \begin{cases} \frac{65}{6}, & p = 1, \\ [-\frac{20}{3}, 35]^\top, & p = 2, \\ [-\frac{5}{2}, 10, 25]^\top, & p = 3, \\ [10, -140, 400, -250, 0, \ldots, 0]^\top, & p \geqslant 4. \end{cases}$$

Hence, the approximation error $\mathbb{E}\left(g^{\mathcal{G}_p}(\boldsymbol{X}) - g^*(\boldsymbol{X})\right)^2$ is given by

$$\int_0^1 \left([1, u, \ldots, u^{p-1}]\,\boldsymbol{\beta}_p - [1, u, u^2, u^3]\,\boldsymbol{\beta}^*\right)^2 \mathrm{d}u = \begin{cases} \frac{32225}{252} \approx 127.9, & p = 1, \\ \frac{1625}{63} \approx 25.8, & p = 2, \\ \frac{625}{28} \approx 22.3, & p = 3, \\ 0, & p \geqslant 4. \end{cases}$$

In general, as the class of approximating functions $\mathcal{G}$ becomes more complex, the approximation error goes down.

Statistical Error

Since $g_\tau(\boldsymbol{x}) = \boldsymbol{x}^\top \widehat{\boldsymbol{\beta}}$, the statistical error can be written as

$$\int_0^1 \left([1, \ldots, u^{p-1}](\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_p) \right)^2 \mathrm{d}u = (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_p)^\top \mathbf{H}_p (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_p). \quad (2)$$

Note that the statistical error depends on the estimate $\widehat{\boldsymbol{\beta}}$, which in its turn depends on the training set τ.

In general, as the class of approximating functions $\mathcal{G}$ becomes more complex, the statistical error increases.

Generalization Risk

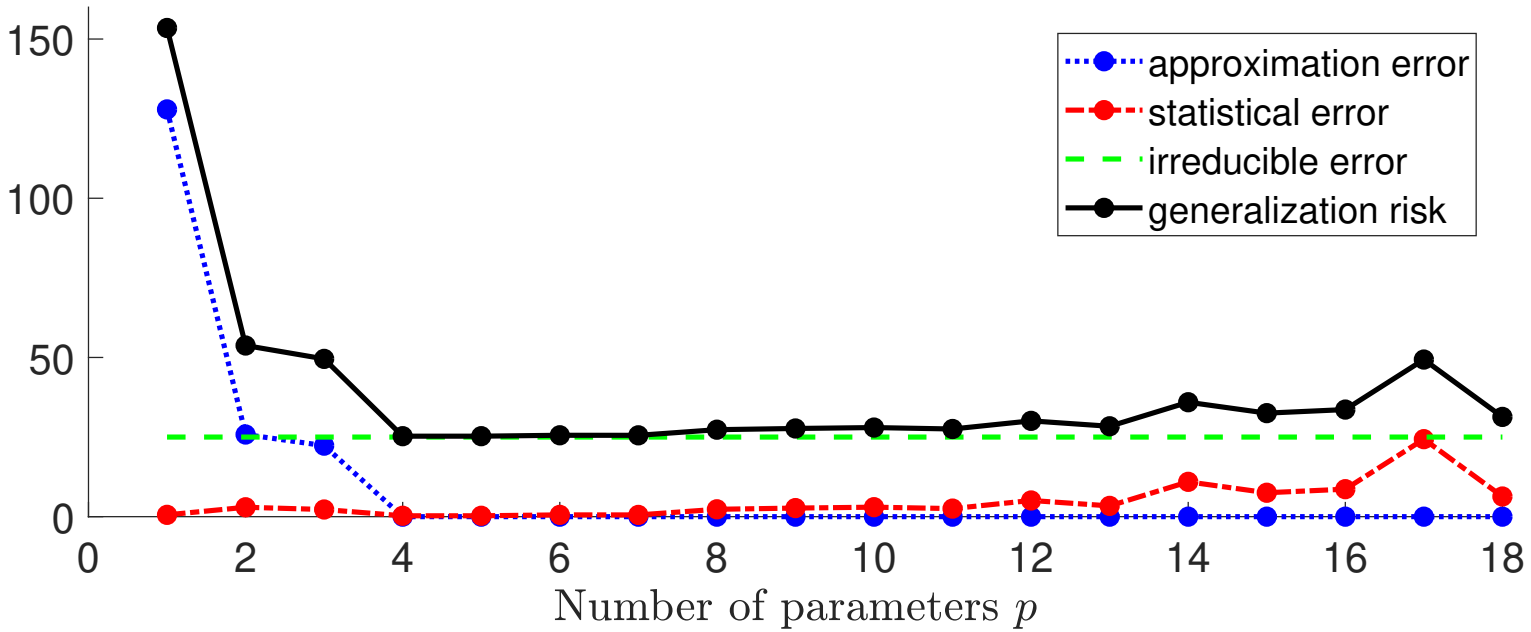


Figure: The generalization risk for a particular training set is the sum of the irreducible error, the approximation error, and the statistical error.

The approximation error decreases to zero as p increases, whereas the statistical error has a tendency to increase after $p = 4$.

Expected Statistical Error

We can obtain a better understanding of the statistical error by considering its expected behavior; that is, averaged over many training sets.

For the squared-error loss (and general $\mathcal{G}$), we can write

$$\ell(g_\tau^{\mathcal{G}}) = \mathbb{E}(g_\tau^{\mathcal{G}}(\boldsymbol{X}) - Y)^2 = \ell^* + \mathbb{E}\left(g_\tau^{\mathcal{G}}(\boldsymbol{X}) - g^*(\boldsymbol{X})\right)^2 = \ell^* + \mathbb{E}D^2(\boldsymbol{X}, \tau),$$

where $D(\boldsymbol{x}, \tau) := g_\tau^{\mathcal{G}}(\boldsymbol{x}) - g^*(\boldsymbol{x})$.

In this decomposition, the statistical error and approximation error are combined.

Bias-Variance Tradeoff

The expection of $D^2(\boldsymbol{x}, \mathcal{T})$ for a random training set $\mathcal{T}$ is:

$$\mathbb{E}\left(g_{\mathcal{T}}^{\mathcal{G}}(\boldsymbol{x}) - g^*(\boldsymbol{x})\right)^2 = \mathbb{E}D^2(\boldsymbol{x}, \mathcal{T}) = (\mathbb{E}D(\boldsymbol{x}, \mathcal{T}))^2 + \mathbb{V}\mathrm{ar}\, D(\boldsymbol{x}, \mathcal{T})$$
$$= \underbrace{(\mathbb{E}g_{\mathcal{T}}^{\mathcal{G}}(\boldsymbol{x}) - g^*(\boldsymbol{x}))^2}_{\text{pointwise squared bias}} + \underbrace{\mathbb{V}\mathrm{ar}\, g_{\mathcal{T}}^{\mathcal{G}}(\boldsymbol{x})}_{\text{pointwise variance}} .$$

- The pointwise squared bias term is a measure for how close $g_{\mathcal{T}}^{\mathcal{G}}(\boldsymbol{x})$ is on average to the true $g^*(\boldsymbol{x})$. It can be reduced by making the class of functions $\mathcal{G}$ *more* complex.
- The pointwise variance measures the squared deviation of $g_{\mathcal{T}}^{\mathcal{G}}(\boldsymbol{x})$ from its expected value $\mathbb{E}g_{\mathcal{T}}^{\mathcal{G}}(\boldsymbol{x})$. it can be reduced by making the class $\mathcal{G}$ *less* complex.

We are thus seeking learners that provide an optimal bias–variance tradeoff.

Expected Generalization Risk

Note that the *expected* generalization risk can be written as

$$\mathbb{E}\,\ell(g_{\mathcal{T}}^{\mathcal{G}}) = \ell^* + \mathbb{E}D^2(\boldsymbol{X}, \mathcal{T}),$$

where $\boldsymbol{X}$ and $\mathcal{T}$ are independent. It therefore decomposes as

$$\mathbb{E}\,\ell(g_{\mathcal{T}}^{\mathcal{G}}) = \ell^* + \underbrace{\mathbb{E}\left(\mathbb{E}[g_{\mathcal{T}}^{\mathcal{G}}(\boldsymbol{X}) \mid \boldsymbol{X}] - g^*(\boldsymbol{X})\right)^2}_{\text{expected squared bias}} + \underbrace{\mathbb{E}[\mathbb{V}\mathrm{ar}[g_{\mathcal{T}}^{\mathcal{G}}(\boldsymbol{X}) \mid \boldsymbol{X}]]}_{\text{expected variance}}.$$