Machine Learning
Linear Regression

Prof. Matthias Hein

Machine Learning Group
Department of Mathematics and Computer Science
Saarland University, Saarbrücken, Germany

Lecture 6, 8.11.2013

Linear methods

Linear methods: most simple regression and classification techniques.

- easy interpretation: feature has a high influence if it has a large weight.
- linear methods: have possibly high bias but low variance $\Rightarrow$ can be fit already with only a few training points.
- often competitive with non-linear methods in high dimensions,
- Using transformations of the input features (**basis functions**) one can easily generate non-linear functions in the input space.

Important: Linear methods are *linear* in the parameters, but not necessarily linear in the original input features.

Least squares regression I

Risk of squared loss:

$$\mathbb{E}[(Y - f(X))^2] = \mathbb{E}\big[\mathbb{E}\big[(Y - f(X))^2 \,|\, X\big]\big].$$

Bayes optimal function:

$$f(x) = \mathbb{E}\big[Y | X = x\big].$$

Definition

Given a training sample $T_n = (X_i, Y_i)_{i=1}^n$ with $X_i \in \mathcal{X}$ and $Y_i \in \mathbb{R}$ and a function space $\mathcal{F}$ we define **least squares regression** as the mapping $\mathcal{A} : T_n \to \mathcal{F}$ with,

$$T_n \mapsto f_n = \operatorname*{arg\,min}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2.$$

Least squares regression II

Linear least squares regression: used **Function class:**

$$\mathcal{F} = \Big\{ f \;\Big|\; f(x) = \sum_{i=1}^{d} w_i x_i + b = \langle w, x \rangle + b, \quad w \in \mathbb{R}^d,\ b \in \mathbb{R} \Big\}.$$

Notation:

- w is the **weight vector**,
- summarize the outputs $(Y_i)_{i=1}^n$ into a column vector $Y \in \mathbb{R}^n$ and the inputs vectors $(X_i)_{i=1}^n$ into a matrix $X \in \mathbb{R}^{n \times d}$,

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \qquad X = \begin{pmatrix} X_{11} & \dots & X_{1d} \\ \vdots & & \vdots \\ X_{n1} & \dots & X_{nd} \end{pmatrix},$$

  where X is called the **design matrix**.
- **Convention in the lecture:** vectors are always column vectors !

Least squares regression III

Constant term in the function class $\mathcal{F}$:

- Add extra dimension to input vector to integrate constant term in the function class,

$$X_i' = (X_{i1}, \ldots, X_{id}, 1) \qquad \text{or} \qquad X_{i(d+1)}' = 1, \ \forall i.$$

An affine function is characterized by the weight vector w,

$$w \in \mathbb{R}^{d+1}, \ f(X_i') = \langle w, X_i' \rangle = \sum_{j=1}^{d+1} w_j X_{ij}' = \sum_{j=1}^{d} w_j X_{ij} + w_{d+1}.$$

Linear least squares regression:

$$w_n = \operatorname*{arg\,min}_{w \in \mathbb{R}^{d+1}} \quad \frac{1}{n} \sum_{i=1}^{n} (Y_i - \langle X_i, w \rangle)^2$$

Convention: we make the constant b explicit in the lecture

Least squares regression IV

Proposition

Let $X \in \mathbb{R}^{n \times d}$. The solution w_n of linear least squares regression is given by

$$w_n = (X^T X)^{-1} X^T Y,$$

where the inverse $(X^T X)^{-1}$ exists if X has rank d. If X has not rank d, then $(X^T X)^{-1}$ has to be understood in the sense of a generalized inverse. In this case the solution is not unique but if w_n^1, w_n^2 are two solutions, then the predictions agree on the training data

$$f_{w_n^1}(X_i) = \langle w_n^1, X_i \rangle = \langle w_n^2, X_i \rangle = f_{w_n^2}(X_i), \quad \textit{for all } i = 1, \ldots, n.$$

Proof: Objective function of the optimization problem with $w \in \mathbb{R}^d$,

$$O_{LLSR}(w) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \langle w, X_i\rangle)^2 = \frac{1}{n}\|Y - Xw\|^2 .$$

Taking the derivative with respect to w,

$$\nabla_w O_{LLSR} = -\frac{2}{n}X^T(Y - Xw).$$

The necessary condition for an extremum of O_{LLSR} is therefore

$$\frac{2}{n}X^T(Y - Xw) = 0 \quad \Longrightarrow \quad X^T Y = (X^T X)w \qquad w_n = (X^T X)^{-1}X^T Y$$

Hessian of the objective function $\frac{2}{n}X^T X \Rightarrow$ positive-definite if X has rank d. If X has rank smaller than d, then w_n defined using the generalized inverse is a solution and every $w = w_n + v$ where v is orthogonal to the subspace $\mathrm{Span}\{X_1, \ldots, X_n\}$ is another solution.

The pseudo-inverse

$(X^T X)^{-1} X^T$ is the Moore-Penrose **pseudo inverse** of X if X has rank d.

Definition

Let $A \in \mathbb{R}^{m \times n}$ with rank $r \leq \min\{m, n\}$. Then the **pseudo-inverse** A^+ of A is defined as

$$A^+ = \underset{B \in \mathbb{R}^{n \times m}}{\arg\min} \left\| AB - \mathbb{1}_m \right\|_F^2 ,$$

where $\|\cdot\|_F$ is the **Frobenius norm** ($\|A\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n A_{ij}^2}$) and $\mathbb{1}_m$ the identity matrix in $\mathbb{R}^m$.

Let A be a square matrix which is invertible, then

$$(A^T A)^{-1} A^T = A^{-1} (A^T)^{-1} A^T = A^{-1}.$$

The pseudo-inverse II

The **singular value decomposition** of $A \in \mathbb{R}^{m\times n}$,

$$A = U \Sigma V^T,$$

- U is an orthogonal matrix $U = (u_1, \ldots, u_m) \in \mathbb{R}^{m\times m}$, that is $U^T U = \mathbb{1}_m$,
- V is an orthogonal matrix $V = (v_1, \ldots, v_n) \in \mathbb{R}^{n\times n}$, that is $V^T V = \mathbb{1}_n$,
- $\Sigma \in \mathbb{R}^{m\times n}$ with $\Sigma_{ij} = \begin{cases} \sigma_i & \text{if } i = j \text{ and } i \leq r, \\ 0 & \text{otherwise} \end{cases}$.

  The $\sigma_i > 0$, $i = 1, \ldots, r$ are the **singular values** of A.

The **pseudo inverse** A^+ is then given by

$$A^+ = V \Sigma^+ U^T,$$

where $\Sigma^+ \in \mathbb{R}^{n\times m}$ is defined as $\Sigma^+_{ij} = \begin{cases} 1/\sigma_i & \text{if } i = j \text{ and } i \leq r, \\ 0 & \text{otherwise} \end{cases}$.

The pseudo-inverse III

The **pseudo inverse** A^+ is then given by

$$A^+ = V\Sigma^+ U^T,$$

where $\Sigma^+ \in \mathbb{R}^{n\times m}$ is given by $\Sigma^+_{ij} = \begin{cases} 1/\sigma_i & \text{if } i=j \text{ and } i \le r, \\ 0 & \text{otherwise} \end{cases}.$

Let $A \in \mathbb{R}^{n\times m}$. Given that $m \le n$ and $\mathrm{ran}(A) = m$, one can write the pseudo inverse A^+ as $A^+ = (A^T A)^{-1} A^T$,

$$\begin{aligned}(A^T A)^{-1} A^T &= (V\Sigma^T U^T U \Sigma V^T)^{-1} V \Sigma^T U^T = (V \Sigma^T \Sigma V^T)^{-1} V \Sigma^T U^T \\ &= V(\Sigma^T\Sigma)^{-1} V^T V \Sigma^T U^T = V(\Sigma^T \Sigma)^{-1} \Sigma^T U^T = V\Sigma^+ U^T.\end{aligned}$$

Basis functions/Feature maps:

- map the input $x \to \phi(x)$,
  $\mathcal{X} = \mathbb{R}: \quad x, x^2, x^3, \ldots$ (polynomials),
  $\mathcal{X} = [0, 2\pi]: \quad \sin(x), \cos(x), \sin(2x), \cos(2x), \ldots$ (Fourier basis).

Fixed, pre-defined set of D **basis functions**, $\phi_i : \mathbb{R}^d \to \mathbb{R}$, we define the function space

$$\mathcal{F} = \Big\{ f : \mathbb{R}^d \to \mathbb{R},\ f(x) = \sum_{i=1}^{D} w_i\, \phi_i(x) \,\Big|\, w \in \mathbb{R}^D \Big\}.$$

Advantage: explicit integration of prior knowledge possible.

Basis functions II

Generalized design matrix: $\Phi \in \mathbb{R}^{n \times D}$,

$$\Phi = \begin{pmatrix} \phi_1(X_1) & \ldots & \phi_D(X_1) \\ \vdots & & \vdots \\ \phi_1(X_n) & \ldots & \phi_D(X_n) \end{pmatrix},$$

Least squares regression problem:

$$w_n = \underset{w \in \mathbb{R}^D}{\arg\min} \, \frac{1}{n} \sum_{i=1}^{n} (Y_i - \langle w, \phi(X_i) \rangle)^2 = \frac{1}{n} \left\| Y - \Phi w \right\|^2,$$

with solution

$$w_n = (\Phi^T \Phi)^{-1} \Phi^T Y,$$

where the matrix $(\Phi^T \Phi)^{-1} \Phi^T \in \mathbb{R}^{D \times n}$ is the pseudo-inverse of Φ.

Basis functions III

Properties:

- The final function, $f(x) = \langle w_n, \phi(x) \rangle = \sum_{i=1}^{D} w_i \phi_i(x)$, is linear in the parameter w,
- allows direct modeling of prior knowledge,
- function space $\mathcal{F} = \left\{ f(x) = \sum_{i=1}^{D} w_i \, \phi_i(x) \,|\, w \in \mathbb{R}^D \right\}$ is D-dimensional,
- Problem: want to model all polynomials in $\mathbb{R}^d$, d polynomials of degree one (linear functions), $\frac{d(d+1)}{2}$ polynomials of degree two, ... Set of basis functions increases rapidly with the dimension $d \Rightarrow$ not practical.

Ridge Regression - Least Squares with L_2-Regularization

Ridge regression:

- Motivation: originally: add small ridge to the solution so that it becomes unique,
  today: regularized version of the least squares problem.
- Function space: $\mathcal{F} = \left\{ f(x) = \sum_{i=1}^{D} w_i \, \phi_i(x) \,|\, w \in \mathbb{R}^D \right\}$
- Loss: squared loss
- Regularizer: $\Omega(w) = \sum_{i=1}^{D} w_i^2 = \|w\|_2^2$.

Definition

Given sample $T_n = (X_i, Y_i)_{i=1}^n$, **ridge regression** is defined as the mapping $\mathcal{A} : T_n \to \mathcal{F}$ with,

$$T_n \mapsto f_n = \mathop{\arg\min}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \langle w, \phi_i(x) \rangle)^2 + \lambda \sum_{i=1}^{D} w_i^2.$$

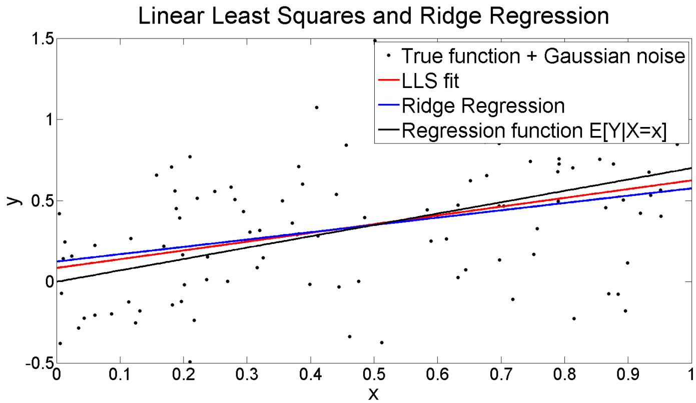


Figure : Linear least squares regression versus linear ridge regression. The regression function is linear.

Result II

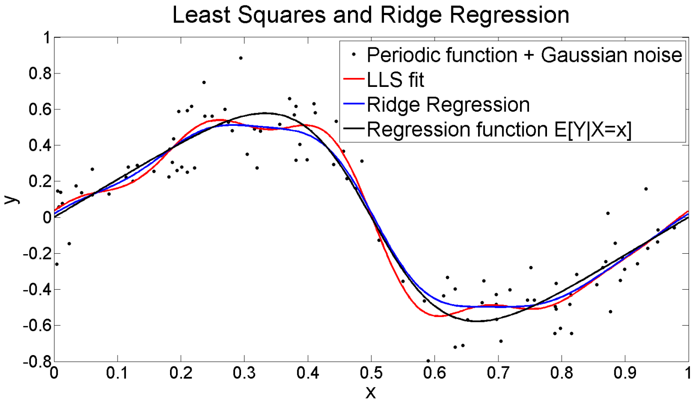


Figure : Comparison of least squares and ridge regression using a set of periodic basis functions.

Ridge Regression - Least Squares with L_2-Regularization II

Solution of ridge regression:

$$w_{n,\lambda} = (\Phi^T \Phi + \lambda \mathbb{1}_D)^{-1} \Phi^T Y.$$

Properties:

- solution $w_{n,\lambda}$ exists and is unique,
- regularizer $\Omega(w) = \|w\|^2$ corresponds to

$$p(w) \propto e^{-\Omega(w)} = e^{-\|w\|^2}.$$

as a prior for maximum a posteriori (MAP) estimation

Geometric interpretation

Linear least squares regression: use SVD of X, $X = U\Sigma V^T$, where rank Σ=r,

$$Xw_n = X(X^TX)^{-1}X^TY = U\Sigma V^T V^T(\Sigma^+)^2 V^T V\Sigma^T U^T Y = \sum_{i=1}^{r} u_i \langle u_i, Y\rangle .$$

Ridge regression:

$$Xw_{n,\lambda} = X(X^TX + \lambda \mathbb{1}_d)^{-1}X^TY = UF(\Sigma)U^TY = \sum_{i=1}^{r} u_i \frac{\sigma_i^2}{\sigma_i^2 + \lambda} \langle u_i, Y\rangle ,$$

where $F(\Sigma) = \begin{cases} \frac{\sigma_i^2}{\sigma_i^2+\lambda} & \text{if } i = j \text{ and } i \leq r, \\ 0 & \text{otherwise} \end{cases}$, σ_i are singular values of X.

- outputs are projected on the basis spanned by U,
- The directions $u_i = \frac{1}{\sigma_i}Xv_i$ correspond to the (mapped) eigenvectors v_i of the covariance matrix $C_{ij} = X^TX$ if X is centered.

Other regularization functionals: $\Omega(w) = \sum_{i=1}^{n} |w_i|^p = \|w\|_p^p$.
$\Rightarrow$ L_2-norm is the only **isotropic** norm in the family of p-norms !

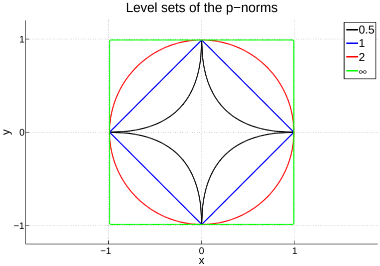


Figure : The level set $\|w\|_p = 1$ of the p-norms. Note that the $\|\cdot\|_p$ is only a norm for $p \geq 1$, in which case the unit-ball is a convex set. Clearly for $p = 0.5$ the "unit-ball" is not convex.

The lasso - Least Squares with L_1-Regularization II

Definition

Given a training sample $T_n = (X_i, Y_i)_{i=1}^n$ with $X_i \in \mathcal{X}$ and $Y_i \in \mathbb{R}$ and the function space $\mathcal{F} = \{\sum_{j=1}^D w_j \phi_j(x) \,|\, w \in \mathbb{R}^D\}$ we define **the lasso** as the mapping $\mathcal{A} : T_n \to \mathcal{F}$ with,

$$T_n \mapsto w_n = \operatorname*{arg\,min}_{w \in \mathbb{R}^D} \frac{1}{n} \sum_{i=1}^n (Y_i - \langle w, \Phi(X_i) \rangle)^2 + \lambda \sum_{i=1}^D |w_i|.$$

Motivation:

- L_1-norm induces **sparsity** (a lot of components w_i of w are zero).
  Why ? The "zero norm" (not really a norm) enforces directly sparsity:
  $$\|w\|_0 = \sum_{i=1}^{D} \mathbb{1}_{w_i \neq 0}.$$
  L_1-norm is the norm which is "closest" to the "zero norm" !
  Sparsity is good : less storage, faster evalution $f(x) = \langle w, x \rangle$, feature selection

The lasso - Least Squares with L_1-Regularization III

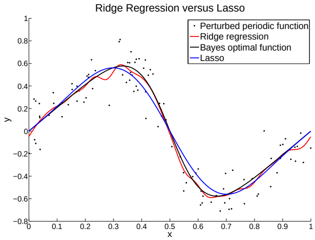
Motivation:

- L_1-norm induces **sparsity** (a lot of components w_i of w are zero).
  Why ? The "zero norm" (not really a norm) enforces directly sparsity:

  $$\|w\|_0 = \sum_{i=1}^{D} \mathbb{1}_{w_i \neq 0}.$$

  L_1-norm is the norm which is "closest" to the "zero norm" !
  Sparsity is good : less storage, faster evalution $f(x) = \langle w, x \rangle$, feature selection
- $\|w\|_2^2$ penalizes large weights heavily $\Rightarrow$ preference for small weights in all directions. (regularizer is **isotropic**)
  $\|w\|_1$ penalizes large and small weights "equally" $\Rightarrow$ produces often large weights in few directions.

Comparison: lasso and ridge regression



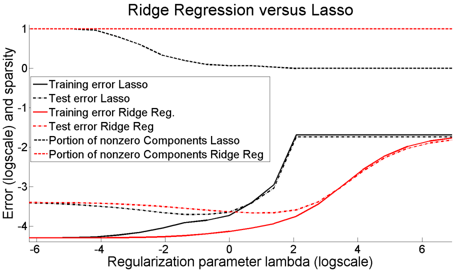


Figure : **Left:** Perturbed training data and regression function in black, we show the solution of ridge regression in blue and of Lasso in red for $\lambda = 1$,
Right: Behavior of training and test error and number of non-zero components of the weight vector as a function of the regularization parameter λ.

Bias and variance of estimators

Solutions w_n of least squares or ridge regression are estimators for the optimal parameter w^* (Bayes optimal **linear** function for the squared loss),

$$w^* = \arg\min_{w \in \mathbb{R}^d} \mathbb{E}\big[(Y - \langle w, X\rangle)^2\big] = \mathbb{E}\big[(Y - \sum_{i=1}^{d} w_i X_i)^2\big],$$

The solution can be derived as (X is a row vector !):

$$w^* = \Big(\mathbb{E}[X^T X]\Big)^{-1} \mathbb{E}[X^T Y].$$

The empirical solutions w_n depend on the training sample $T = (X_i, Y_i)$.

Questions:

- Is the average estimator w_n over training samples of size n equal to the optimal w^* ?
- How much does the estimator w_n fluctuate around its average value over all possible training samples from P of size n ?

Bias and variance of estimators II

Definition

Given a sample $T = (X_i)_{i=1}^n$ and an estimate (also called statistics) $f_n : T \to \mathbb{R}$ of a quantity $f \in \mathbb{R}$ the **bias** of f_n is defined as

$$\operatorname{Bias} f_n = \mathbb{E}_T\big[f_n\big] - f,$$

the difference of the expectation of f_n over all training sets T (all possible i.i.d. training sets of size n) and the true quantity f.

- The estimator f_n is said to be **unbiased** if the bias is zero.
- It is **asymptotically unbiased** if $\lim_{n\to\infty} \operatorname{Bias} f_n = 0$.

The **variance** of f_n is defined as,

$$\operatorname{Var} f_n = \mathbb{E}_T\big[(f_n - \mathbb{E}_T[f_n])^2\big].$$

Bias and variance of estimators II

Examples for bias and variance:

- **The empirical mean** $\mathbb{E}_{\mathrm{P}_n}[X] = \frac{1}{n}\sum_{i=1}^n X_i$ is an **estimator of the true mean** $\mathbb{E}[X] = \mathbb{E}_{\mathrm{P}}[X]$.

$$\mathbb{E}_T\Big[\frac{1}{n}\sum_{i=1}^n X_i\Big] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{X_i}\big[X_i\big] = \frac{1}{n} n\mathbb{E}[X] = \mathbb{E}[X] \implies \text{unbiased}\,!$$

- **empirical variance** $\mathrm{Var}_{\mathrm{P}_n}[X] = \frac{1}{n}\sum_{i=1}^n (X_i - \mathbb{E}_{\mathrm{P}_n}[X])^2$ as an estimator of the **true variance** $\mathrm{Var}_{\mathrm{P}}[X] = \mathrm{Var}[X]$.

$$\mathbb{E}_T\big[\,\mathrm{Var}_{\mathrm{P}_n}[X]\big] = \frac{n-1}{n}\mathrm{Var}[X] \implies \text{biased}\,!\ \text{underestimation}\,!$$

The estimator $\frac{1}{n-1}\sum_{i=1}^n (X_i - \mathbb{E}_{\mathrm{P}_n}[X])^2$ for the variance of X is unbiased.

Bias and variance of estimators III

The risk $R(f_n)$, the expected squared loss, of the estimator f_n:

$$
\begin{aligned}
R(f_n) &= \mathbb{E}[(Y - f_n(X))^2] = \mathbb{E}\big[\,\mathbb{E}[(Y - f_n(X))^2|X]\big]\\
&= \mathbb{E}\big[\,\mathbb{E}[(Y - \mathbb{E}[Y|X] + \mathbb{E}[Y|X] - f_n(X))^2|X]\,\big]\\
&= \mathbb{E}\big[\,\mathbb{E}[(Y - \mathbb{E}[Y|X])^2|X]\,\big] + \mathbb{E}\big[\,\mathbb{E}[(\mathbb{E}[Y|X] - f_n(X))^2|X]\,\big]\\
&\quad + 2\mathbb{E}\big[\,\mathbb{E}[(Y - \mathbb{E}[Y|X])(\mathbb{E}[Y|X] - f_n(X))|X]\,\big],\\
&= \mathbb{E}\big[(Y - \mathbb{E}[Y|X])^2\big] + \mathbb{E}\big[(\mathbb{E}[Y|X] - f_n(X))^2\big]
\end{aligned}
$$

Interpretation:

- The first term is the **Bayes optimal risk** (often also called noise term),
  $\eta(x) = \mathbb{E}[Y|X = x]$ is the Bayes optimal function for the squared loss.
- The second term measures the **deviation of f_n from the Bayes optimal function**. It is a random quantity since f_n depends on the training data !

Expected risk $\mathbb{E}_T[R(f_n)]$ over all possible training sets T:

$$\mathbb{E}_T\big[R(f_n)\big] = \mathbb{E}\big[(Y-\eta(X))^2\big] + \mathbb{E}_T\big[\mathbb{E}_X[(\eta(X)-f_n(X))^2]\big],$$

The first term is constant !

$$\begin{aligned}
\mathbb{E}_T\big[(f_n(x)-\eta(x))^2\big] &= \mathbb{E}_T\Big[\Big(f_n(x)-\mathbb{E}_T f_n(x)+\mathbb{E}_T f_n(x)-\eta(x)\Big)^2\Big] \\
&= \mathbb{E}_T\Big[\Big(f_n(x)-\mathbb{E}_T[f_n(x)]\Big)^2\Big] + \mathbb{E}_T\Big[\Big(\mathbb{E}_T[f_n(x)]-\eta(x)\Big) \\
&\quad + 2\mathbb{E}_T\Big[\Big(f_n(x)-\mathbb{E}_T f_n(x)\Big)\Big(\mathbb{E}_T f_n(x)-\eta(x)\Big)\Big] \\
&= \mathbb{E}_T\Big[\Big(f_n(x)-\mathbb{E}_T[f_n(x)]\Big)^2\Big] + \Big(\mathbb{E}_T[f_n(x)]-\eta(x)\Big)^2 \\
&= \mathrm{Var}\, f_n(x) + (\mathrm{Bias}\, f_n(x))^2,
\end{aligned}$$

(Noise)-Bias-Variance-Decomposition:

$$\mathbb{E}_T\big[R(f_n)\big] = \mathbb{E}\big[(Y-\eta(X))^2\big] + \mathbb{E}\Big[\big(\mathrm{Bias}\, f_n(X)\big)^2\Big] + \mathbb{E}\big[\mathrm{Var}\, f_n(X)\big],$$

where

- **Noise term at x:** $\mathbb{E}\big[(Y-\eta(X))^2 | X = x\big]$,
- **Variance of f_n:** $\mathrm{Var}\, f_n(x) = \mathbb{E}_T\big[(f_n(x) - \mathbb{E}_T[f_n(x)])^2\big]$,
- **Bias of f_n:** $\mathrm{Bias}\, f_n(x) = \mathbb{E}_T[f_n(x)] - \eta(x)$,

(Noise)-Bias-Variance-Decomposition:

$$\mathbb{E}_T\big[R(f_n)\big] = \mathbb{E}\big[(Y-\eta(X))^2\big] + \mathbb{E}\Big[\big(\operatorname{Bias} f_n(X)\big)^2\Big] + \mathbb{E}\big[\operatorname{Var} f_n(X)\big],$$

where

- **Noise term at x:** $\mathbb{E}\big[(Y-\eta(X))^2 | X = x\big]$,
- **Variance of f_n:** $\operatorname{Var} f_n(x) = \mathbb{E}_T\big[(f_n(x) - \mathbb{E}_T[f_n(x)])^2\big]$,
- **Bias of f_n:** $\operatorname{Bias} f_n(x) = \mathbb{E}_T[f_n(x)] - \eta(x)$,

$$\boxed{\text{expected loss} = \text{noise} + \text{variance} + \text{squared bias.}}$$

Trade-off between **bias** and **variance**
corresponds to
Trade-off between **overfitting** and **underfitting**.

Bias-Variance of Least Squares

Bias-Variance-Decomposition for the Least-Squares estimator:

$f_n = \langle w_n, x \rangle \Rightarrow$ express bias and variance of f_n via the bias and covariance of w_n,

$$\begin{aligned}
\mathrm{Bias}\, f_n(x) &= \mathbb{E}_T[f_n(x)] - f^*(x) = \mathbb{E}_T[\langle w_n, x \rangle] - \langle w^*, x \rangle \\
&= \langle \mathbb{E}_T[w_n] - w^*, x \rangle \\
&= \langle \mathrm{Bias}\, w_n, x \rangle \,, \\
\mathrm{Var}\, f_n(x) &= \mathbb{E}_T\Big[\big(f_n(x) - \mathbb{E}_T[f_n(x)]\big)^2\Big] = \mathbb{E}_T\Big[\big(\langle w_n, x \rangle - \langle \mathbb{E}_T[w_n], x \rangle\big)^2\Big] \\
&= \mathbb{E}_T\Big[\, \langle w_n - \mathbb{E}_T[w_n], x \rangle^2 \,\Big] \\
&= \mathbb{E}_T\Big[\, \sum_{i,j=1}^{d} \big((w_n)_i - \mathbb{E}_T[(w_n)_i]\big) x_i x_j \big((w_n)_j - \mathbb{E}_T[(w_n)_j]\big) \\
&= \mathrm{tr}(x x^T \,\mathrm{Cov}\, w_n) = \langle x, (\mathrm{Cov}\, w_n) x \rangle \,,
\end{aligned}$$

Gauss-Markov-Theorem

Theorem (Gauss Markov theorem)

Suppose that the data obeys the linear model

$$Y = \langle w, X \rangle + \epsilon,$$

with $\mathbb{E}[\epsilon | X = x] = 0$, $\mathrm{Var}[\epsilon | X = x] = \sigma^2$ *and errors at different points are uncorrelated.*
Then

- *the least squares estimator* $w_n = (X^T X)^{-1} X^T Y$ *is unbiased,*
- *among all possible unbiased estimators of the weight vector* w *it has the smallest variance.*

Gauss-Markov-Theorem II

The Gauss-Markov-Theorem is only of very limited practical use:

- Model assumption has to be true !
  In reality linearity is not often encountered
- If the model assumption is correct:
  least squares estimator is the best among all possible **unbiased** estimators !
  $\Rightarrow$ a slightly biased estimator (e.g. ridge regression or lasso) could have much smaller variance
  $\Rightarrow$ better **expected squared error** $\Rightarrow$ **Better estimator !**