

# Probabilistic model

So far in this course, a *model* referred to a parametric family of functions, such as

$$f(x) = \beta_0 + \beta_1 x.$$

We will now discuss *probabilistic (generative) models*, such as

$$\begin{aligned} y &= \beta_0 + \beta_1 x + \epsilon, \\ \epsilon &\sim \mathcal{N}(0, \sigma^2), \end{aligned}$$

where  $\mathcal{N}(0, \sigma^2)$  denotes a Gaussian distribution with mean 0 and variance  $\sigma^2$ .



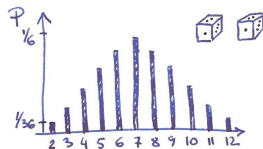
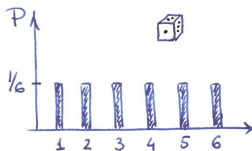
# Probability theory recap

---

# Discrete probability distributions

Probability distributions can be *discrete* or *continuous*.

A discrete random variable  $X$  is described by a *probability mass function* (PMF):



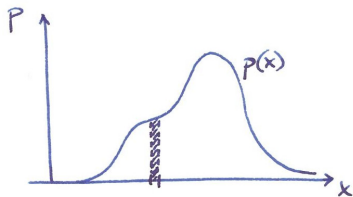
Here  $X$  can take values  $x_i$  with probabilities  $p_i$ , with

$$p_i \geq 0, \quad \sum_i p_i = 1.$$



# Continuous probability distributions

A continuous random variable  $X$  is described by a *probability density function* (PDF):



$$p(x) \geq 0,$$

$$\int_{\mathbb{R}} p(x) dx = 1.$$



# The mean and the variance

If a discrete random variable  $X$  takes values  $x_i$  with probabilities  $p_i$ , then

$$\mathbb{E}[X] = \sum_i x_i p_i,$$

$$\text{Var}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \sum_i (x_i - \mathbb{E}[X])^2 p_i.$$

If a continuous random variable  $X$  is described by a PDF  $p(x)$ , then

$$\mathbb{E}[X] = \int x p(x) dx,$$

$$\text{Var}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \int (x - \mathbb{E}[X])^2 p(x) dx.$$



# Variance, covariance, and correlation

We defined

$$\text{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2].$$

We can similarly define

$$\text{Cov}[X, Y] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])].$$

Note that

$$\text{Cov}[X, X] = \text{Var}[X].$$

If  $\text{Cov}[X, Y] = 0$ , then  $X$  and  $Y$  are *uncorrelated*. Reminder:

$$\text{Corr}[X, Y] = \frac{\text{Cov}[X, Y]}{\sqrt{\text{Var}[X] \text{Var}[Y]}}.$$



# Some properties of mean and variance

Some useful properties of the expected value:

$$\mathbb{E}[aX] = a\mathbb{E}[X]$$

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

$$\mathbb{E}[XY] \neq \mathbb{E}[X]\mathbb{E}[Y] \text{ (unless independent)}$$

and of variance:

$$\text{Var}[aX] = a^2 \text{Var}[X]$$

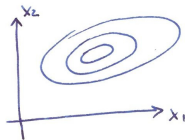
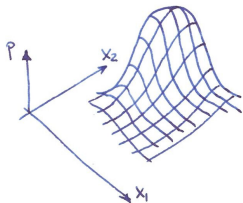
$$\text{Var}[X + Y] \neq \text{Var}[X] + \text{Var}[Y] \text{ (unless uncorrelated)}$$

$$\text{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$



# Multivariate probability distributions

A random variable  $X$  can be *multivariate* (random vector):



$$p(\mathbf{x}) \geq 0,$$

$$\int_{\mathbb{R}^2} p(\mathbf{x}) d\mathbf{x} = 1.$$





# Multivariate probability distributions

If a continuous multivariate random variable  $X$  is described by a PDF  $p(\mathbf{x})$ , then

$$\mathbb{E}[X] = \int \mathbf{x}p(\mathbf{x})d\mathbf{x}.$$

The variance is now replaced by a *covariance matrix*:

$$\text{Cov}[X] = \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^\top].$$

Its diagonal elements are variances of  $X_i$ :

$$\text{Cov}[X]_{ii} = \text{Var}[X_i]$$

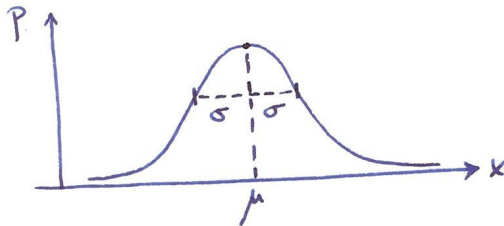
while off-diagonal elements are covariances of  $X_i$  and  $X_j$ :

$$\text{Cov}[X]_{ij} = \text{Cov}[X_i, X_j].$$



# Gaussian distribution

Gaussian (normal) distribution  $\mathcal{N}(\mu, \sigma^2)$  with mean  $\mu$  and variance  $\sigma^2$ :



$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp \left[ -\frac{(x - \mu)^2}{2\sigma^2} \right]$$

If  $\mu = 0$  and  $\sigma = 1$ , this is called *standard* normal distribution:

$$p(x) = \frac{1}{\sqrt{2\pi}} \exp \left[ -\frac{x^2}{2} \right].$$



# Multivariate Gaussian distribution

Multivariate Gaussian distribution  $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  in  $\mathbb{R}^k$  with mean  $\boldsymbol{\mu}$  and covariance matrix  $\boldsymbol{\Sigma}$ :

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^k \det(\boldsymbol{\Sigma})}} \exp \left[ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \right].$$

If  $\boldsymbol{\mu} = \mathbf{0}$  and  $\boldsymbol{\Sigma} = \mathbf{I}$ , this is also called *standard* multivariate normal distribution:

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^k}} \exp \left[ -\frac{1}{2}\|\mathbf{x}\|^2 \right].$$



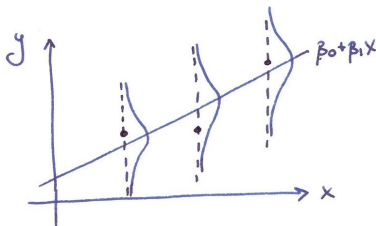
# Back to the probabilistic model for linear regression

---

# Probabilistic model

Probabilistic model for regression:

$$y = \beta_0 + \beta_1 x + \epsilon = \boldsymbol{\beta}^\top \mathbf{x} + \epsilon,$$
$$\epsilon \sim \mathcal{N}(0, \sigma^2).$$



Note: this assumes uncorrelated noise (errors) and equal noise variance for all points (*homoscedasticity*).



# Likelihood

For a given  $\beta$  and given  $\mathbf{x}_i$ ,

$$y \sim \mathcal{N}(\beta^\top \mathbf{x}_i, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp \left[ -\frac{(y - \beta^\top \mathbf{x}_i)^2}{2\sigma^2} \right].$$

Probability density to generate the entire training set  $\{(\mathbf{x}_i, y_i)\}$  is

$$\prod_i \frac{1}{\sigma\sqrt{2\pi}} \exp \left[ -\frac{(y_i - \beta^\top \mathbf{x}_i)^2}{2\sigma^2} \right].$$

If we re-interpret this as a function of  $\beta$  (and  $\sigma^2$ ), then it is called *the likelihood*.



# Maximum likelihood

Find  $\beta$  and  $\sigma^2$  maximizing the likelihood:

$$\prod_i \frac{1}{\sigma\sqrt{2\pi}} \exp \left[ - \frac{(y_i - \beta^\top \mathbf{x}_i)^2}{2\sigma^2} \right].$$

Product of exponentials is annoying to work with  $\Rightarrow$  take the logarithm to obtain *log-likelihood*:

$$\begin{aligned} \sum_i \left[ \log \left[ \frac{1}{\sigma\sqrt{2\pi}} \right] + \left[ - \frac{(y_i - \beta^\top \mathbf{x}_i)^2}{2\sigma^2} \right] \right] &= \\ &= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_i (y_i - \beta^\top \mathbf{x}_i)^2. \end{aligned}$$

It is often convenient to think about minimizing *the negative log-likelihood*.



# Maximum likelihood

Negative log-likelihood:

$$\begin{aligned}\frac{n}{2} \log(2\pi\sigma^2) + \frac{1}{2\sigma^2} \sum_i (y_i - \beta^\top \mathbf{x}_i)^2 &= \\ &= \frac{n}{2} \log(2\pi\sigma^2) + \frac{1}{2\sigma^2} \|\mathbf{y} - \mathbf{X}\beta\|^2.\end{aligned}$$

Maximizing likelihood is equivalent to minimizing squared error!

Exercise: what is the maximum likelihood solution for  $\sigma^2$ ?





# Statistical properties of $\hat{\beta}$

$\hat{\beta}$  is an estimator of  $\beta$ . It is a random variable that depends on the input data. We are interested in the expected value and the (co)variance of this estimator.

We assume  $\mathbf{X}$  is fixed and  $\beta$  is fixed. The response vector  $\mathbf{y} = \mathbf{X}\beta + \epsilon$  is random. We want to study  $\mathbb{E}[\hat{\beta}]$  and  $\text{Cov}[\hat{\beta}]$  over  $\epsilon$ .



# $\hat{\beta}$ is an unbiased estimator

Theorem:  $\mathbb{E}[\hat{\beta}] = \beta$ , i.e. it is an *unbiased* estimator.

Proof:

$$\begin{aligned}\mathbb{E}[\hat{\beta}] &= \mathbb{E}[(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}] \\ &= \mathbb{E}[(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{X}\beta + \epsilon)] \\ &= \mathbb{E}[(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X}\beta] + \mathbb{E}[(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \epsilon] \\ &= \beta + \mathbf{0} = \\ &= \beta.\end{aligned}$$

Note: here we assumed that  $n > p$  and  $\mathbf{X}$  has *full rank*, i.e. all singular values are non-zero, i.e.  $(\mathbf{X}^\top \mathbf{X})^{-1}$  exists.



# Covariance matrix of $\hat{\beta}$ and Gauss-Markov

Exercise:  $\text{Cov}[\hat{\beta}] = \sigma^2(\mathbf{X}^\top \mathbf{X})^{-1}$ .

$\text{Cov}[\hat{\beta}]$  describes uncertainty around  $\hat{\beta}$ . We see that small singular values of  $\mathbf{X}$  lead to large uncertainty.

**Gauss-Markov theorem:**  $\hat{\beta}$  has the smallest variance among all unbiased linear estimators. It is the *best linear unbiased estimator* (BLUE).

What does this mean exactly? That  $\text{Var}[\mathbf{a}^\top \hat{\beta}] \leq \text{Var}[\mathbf{a}^\top \tilde{\beta}]$  for any vector  $\mathbf{a}$  (or, equivalently,  $\text{Cov}[\tilde{\beta}] - \text{Cov}[\hat{\beta}]$  is a *positive semi-definite matrix*, i.e. all singular values are  $\geq 0$ ).



Is the best linear unbiased estimator  
always the best estimator?

No.



# Underfitting, overfitting, and the bias–variance tradeoff

---

# Polynomial regression

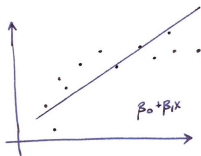
What if we model the relationship between  $y$  and  $x$  but include  $x^2$ ,  $x^3$ , etc. terms?

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3.$$

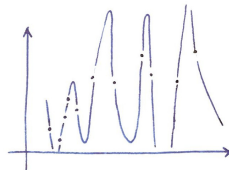
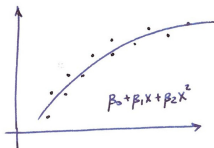
This is *still* linear regression! What?! Yes.



# Underfitting and overfitting



Underfitting  
Model too simple  
High bias



Overfitting  
Model too flexible  
High variance



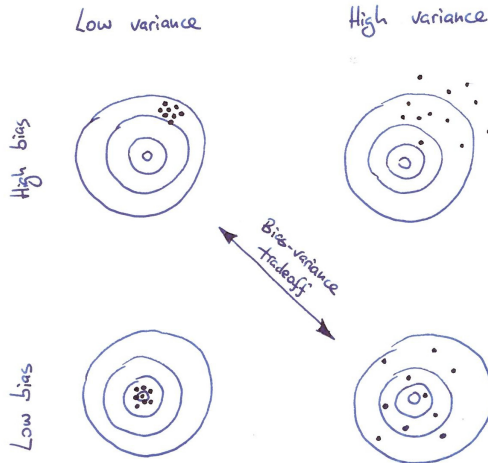
# Bias–variance tradeoff

$$\begin{aligned}\text{MSE} &= \mathbb{E}\left[(y - \hat{f}(x))^2\right] = \\&= \mathbb{E}\left[(f(x) + \epsilon - \hat{f}(x))^2\right] = \\&= \mathbb{E}\left[(f(x) - \hat{f}(x))^2\right] + \sigma^2 = \\&= \mathbb{E}\left[(f(x) - \mathbb{E}[\hat{f}(x)] + \mathbb{E}[\hat{f}(x)] - \hat{f}(x))^2\right] + \sigma^2 = \\&= \left(f(x) - \mathbb{E}[\hat{f}(x)]\right)^2 + \mathbb{E}\left[(\hat{f}(x) - \mathbb{E}[\hat{f}(x)])^2\right] + \\&\quad + 2\left(f(x) - \mathbb{E}[\hat{f}(x)]\right)\mathbb{E}\left[\hat{f}(x) - \mathbb{E}[\hat{f}(x)]\right] + \sigma^2 = \\&= \underbrace{\left(f(x) - \mathbb{E}[\hat{f}(x)]\right)^2}_{\text{Bias}^2} + \underbrace{\mathbb{E}\left[(\hat{f}(x) - \mathbb{E}[\hat{f}(x)])^2\right]}_{\text{Variance}} + \sigma^2 = \\&= \text{Bias}^2 + \text{Variance} + \sigma^2.\end{aligned}$$

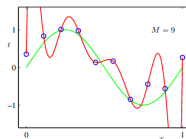
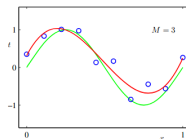
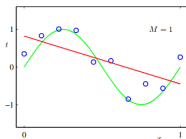
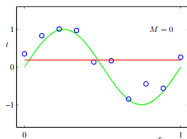




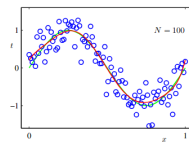
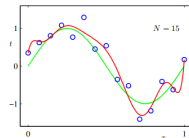
# Intuition for bias and variance



# Overfitting and high variance demonstration



	$M = 0$	$M = 1$	$M = 3$	$M = 9$
$w_0^*$	0.19	0.82	0.31	0.35
$w_1^*$		-1.27	7.99	232.37
$w_2^*$			-25.43	-5321.83
$w_3^*$			17.37	48568.31
$w_4^*$				-231639.30
$w_5^*$				640042.26
$w_6^*$				-1061800.52
$w_7^*$				1042400.18
$w_8^*$				-557682.99
$w_9^*$				125201.43



Bishop, *Pattern Recognition and Machine Learning*



# Training and test error

