



Introduction to Statistical Machine Learning

Cheng Soon Ong & Christian Walder

Machine Learning Research Group

Data61 | CSIRO

and

College of Engineering and Computer Science

The Australian National University

Canberra

February – June 2017

(Many figures from C. M. Bishop, "Pattern Recognition and Machine Learning")



Part II

Introduction

Polynomial Curve Fitting

Probability Theory

Probability Densities

*Expectations and
Covariances*



- Formalise intuitions about problems
- Use language of mathematics to express models
- Geometry, vectors, linear algebra for reasoning
- Probabilistic models to capture uncertainty
- Design and analysis of algorithms
- Numerical algorithms in python
- Understand the choices when designing machine learning methods

What is Machine Learning?

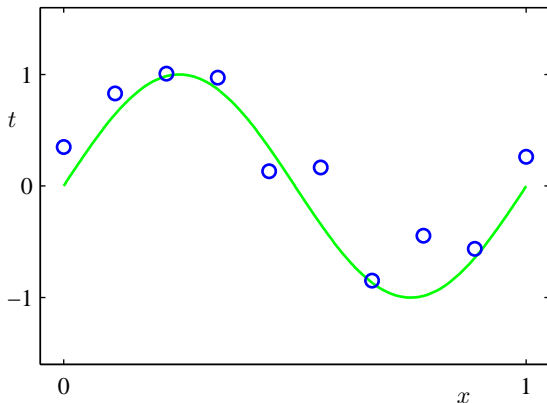


Definition (Mitchell, 1998)

A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P , if its performance at tasks in T , as measured by P , improves with experience E .

- some artificial data created from the function

$$\sin(2\pi x) + \text{random noise} \quad x = 0, \dots, 1$$



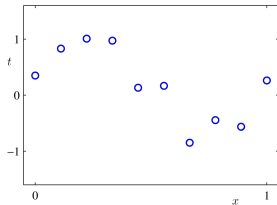
Polynomial Curve Fitting - Input Specification



$$N = 10$$

$$\mathbf{x} \equiv (x_1, \dots, x_N)^T$$

$$\mathbf{t} \equiv (t_1, \dots, t_N)^T$$



Polynomial Curve Fitting - Input Specification



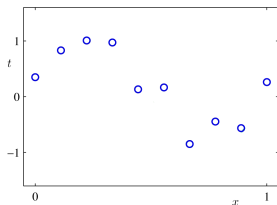
$$N = 10$$

$$\mathbf{x} \equiv (x_1, \dots, x_N)^T$$

$$\mathbf{t} \equiv (t_1, \dots, t_N)^T$$

$$x_i \in \mathbb{R} \quad i = 1, \dots, N$$

$$t_i \in \mathbb{R} \quad i = 1, \dots, N$$

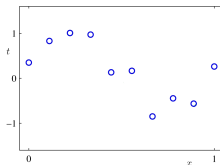


Polynomial Curve Fitting - Model Specification



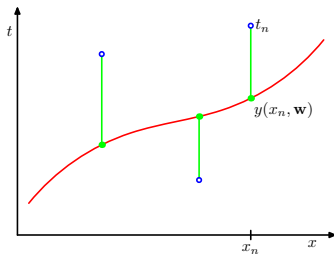
M : order of polynomial

$$\begin{aligned}y(x, \mathbf{w}) &= w_0 + w_1 x + w_2 x^2 + \cdots + w_M x^M \\ &= \sum_{m=0}^M w_m x^m\end{aligned}$$



- nonlinear function of x
- **linear** function of the unknown model parameter \mathbf{w}
- How can we find good parameters $\mathbf{w} = (w_1, \dots, w_M)^T$?

Learning is Improving Performance



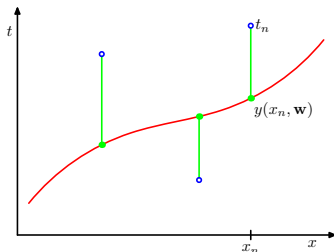
Polynomial Curve Fitting

Probability Theory

Probability Densities

Expectations and
Covariances

Learning is Improving Performance



- Performance measure : Error between target and prediction of the model for the training data

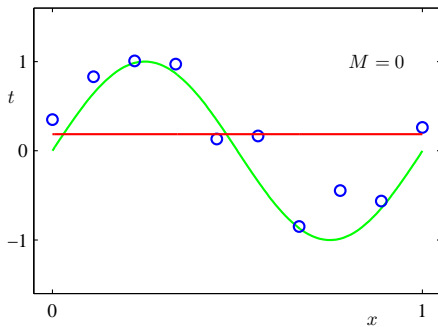
$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N (y(x_n, \mathbf{w}) - t_n)^2$$

- unique minimum of $E(\mathbf{w})$ for argument \mathbf{w}^*

Model Comparison or Model Selection



$$y(x, \mathbf{w}) = \sum_{m=0}^M w_m x^m \quad \Bigg| \quad M=0$$
$$= w_0$$



Polynomial Curve Fitting

Probability Theory

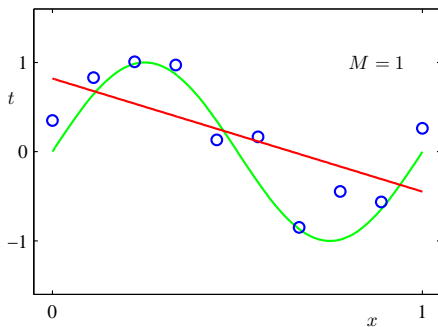
Probability Densities

Expectations and
Covariances

Model Comparison or Model Selection



$$y(x, \mathbf{w}) = \sum_{m=0}^M w_m x^m \quad \Bigg| \quad M=1$$
$$= w_0 + w_1 x$$



Polynomial Curve Fitting

Probability Theory

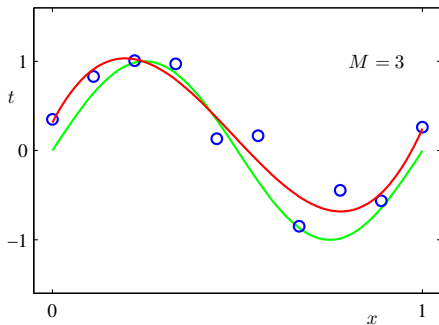
Probability Densities

Expectations and
Covariances

Model Comparison or Model Selection



$$y(x, \mathbf{w}) = \sum_{m=0}^M w_m x^m \quad \Big| \quad M=3$$
$$= w_0 + w_1 x + w_2 x^2 + w_3 x^3$$



Polynomial Curve Fitting

Probability Theory

Probability Densities

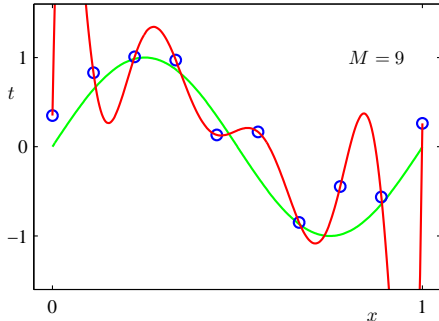
Expectations and
Covariances

Model Comparison or Model Selection



$$y(x, \mathbf{w}) = \sum_{m=0}^M w_m x^m \quad \Big|_{M=9}$$
$$= w_0 + w_1 x + \cdots + w_8 x^8 + w_9 x^9$$

- overfitting



Polynomial Curve Fitting

Probability Theory

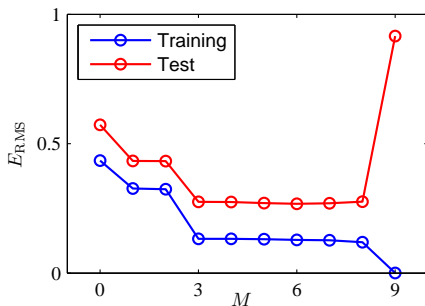
Probability Densities

Expectations and
Covariances

Testing the Model

- Train the model and get \mathbf{w}^*
- Get 100 new data points
- Root-mean-square (RMS) error

$$E_{\text{RMS}} = \sqrt{2E(\mathbf{w}^*)/N}$$



Testing the Model

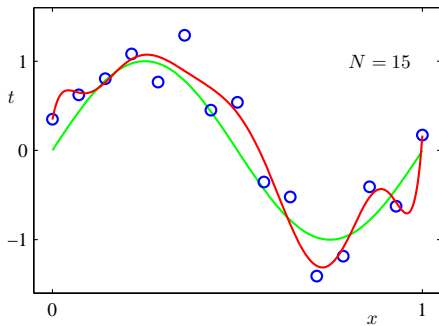


| | M = 0 | M = 1 | M = 3 | M = 9 |
|---------|-------|-------|--------|-------------|
| w_0^* | 0.19 | 0.82 | 0.31 | 0.35 |
| w_1^* | | -1.27 | 7.99 | 232.37 |
| w_2^* | | | -25.43 | -5321.83 |
| w_3^* | | | 17.37 | 48568.31 |
| w_4^* | | | | -231639.30 |
| w_5^* | | | | 640042.26 |
| w_6^* | | | |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

Table: Coefficients w^* for polynomials of various order.



- $N = 15$



Polynomial Curve Fitting

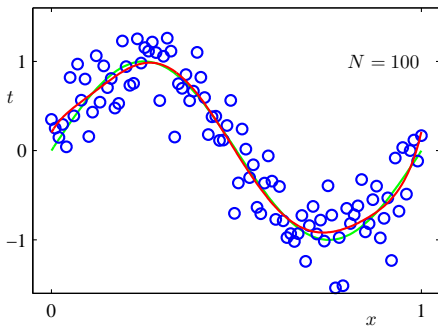
Probability Theory

Probability Densities

*Expectations and
Covariances*



- $N = 100$
- heuristics : have no less than 5 to 10 times as many data points than parameters
- but number of parameters is not necessarily the most appropriate measure of model complexity !
- later: Bayesian approach





- How to constrain the growing of the coefficients \mathbf{w} ?
- Add a **regularisation** term to the error function

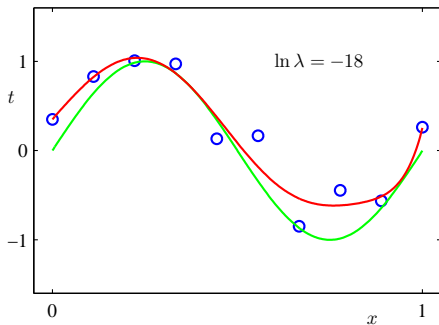
$$\tilde{E}(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N (y(x_n, \mathbf{w}) - t_n)^2 + \frac{\lambda}{2} \|\mathbf{w}\|^2$$

- Squared norm of the parameter vector \mathbf{w}

$$\|\mathbf{w}\|^2 \equiv \mathbf{w}^T \mathbf{w} = w_0^2 + w_1^2 + \dots + w_M^2$$



- $M = 9$



Polynomial Curve Fitting

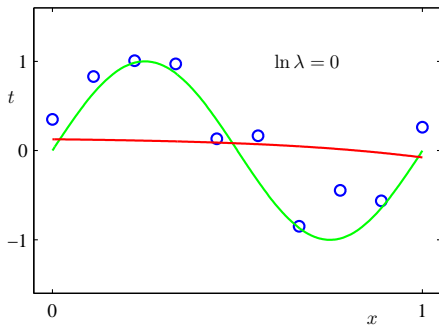
Probability Theory

Probability Densities

Expectations and
Covariances



- $M = 9$



Polynomial Curve Fitting

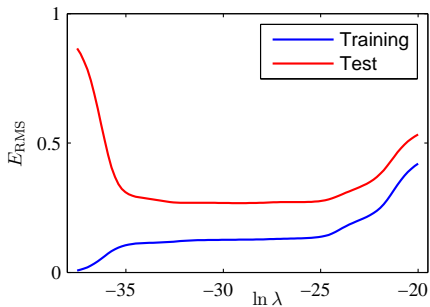
Probability Theory

Probability Densities

*Expectations and
Covariances*



- $M = 9$



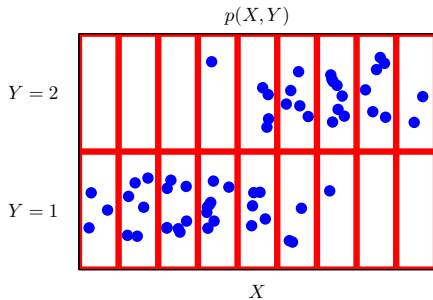
What is Machine Learning?



Definition (Mitchell, 1998)

A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P , if its performance at tasks in T , as measured by P , improves with experience E .

- Task: regression
- Experience: x input examples, t output labels
- Performance: squared error
- Model choice
- Regularisation
- **do not train on the test set!**



Polynomial Curve Fitting

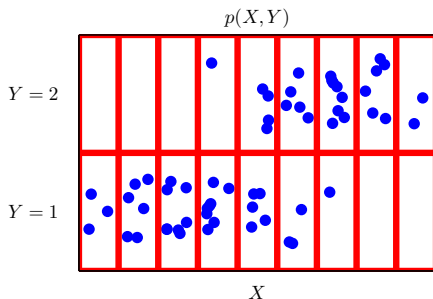
Probability Theory

Probability Densities

Expectations and
Covariances



| Y vs. X | a | b | c | d | e | f | g | h | i | sum |
|---------|---|---|---|---|---|---|---|---|---|-----|
| 2 | 0 | 0 | 0 | 1 | 4 | 5 | 8 | 6 | 2 | 26 |
| 1 | 3 | 6 | 8 | 8 | 5 | 3 | 1 | 0 | 0 | 34 |
| sum | 3 | 6 | 8 | 9 | 9 | 8 | 9 | 6 | 2 | 60 |



Sum Rule



| Y vs. X | a | b | c | d | e | f | g | h | i | sum |
|---------|---|---|---|---|---|---|---|---|---|-----|
| 2 | 0 | 0 | 0 | 1 | 4 | 5 | 8 | 6 | 2 | 26 |
| 1 | 3 | 6 | 8 | 8 | 5 | 3 | 1 | 0 | 0 | 34 |
| sum | 3 | 6 | 8 | 9 | 9 | 8 | 9 | 6 | 2 | 60 |

$$p(X = d, Y = 1) = 8/60$$

$$\begin{aligned} p(X = d) &= p(X = d, Y = 2) + p(X = d, Y = 1) \\ &= 1/60 + 8/60 \end{aligned}$$

$$p(X = d) = \sum_Y p(X = d, Y)$$

$$p(X) = \sum_Y p(X, Y)$$

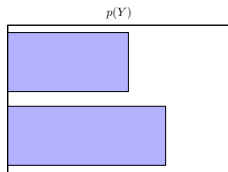
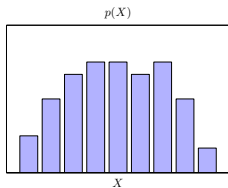
Sum Rule



| Y vs. X | a | b | c | d | e | f | g | h | i | sum |
|---------|---|---|---|---|---|---|---|---|---|-----|
| 2 | 0 | 0 | 0 | 1 | 4 | 5 | 8 | 6 | 2 | 26 |
| 1 | 3 | 6 | 8 | 8 | 5 | 3 | 1 | 0 | 0 | 34 |
| sum | 3 | 6 | 8 | 9 | 9 | 8 | 9 | 6 | 2 | 60 |

$$p(X) = \sum_Y p(X, Y)$$

$$p(Y) = \sum_X p(X, Y)$$





| Y vs. X | a | b | c | d | e | f | g | h | i | sum |
|---------|---|---|---|---|---|---|---|---|---|-----|
| 2 | 0 | 0 | 0 | 1 | 4 | 5 | 8 | 6 | 2 | 26 |
| 1 | 3 | 6 | 8 | 8 | 5 | 3 | 1 | 0 | 0 | 34 |
| sum | 3 | 6 | 8 | 9 | 9 | 8 | 9 | 6 | 2 | 60 |

Conditional Probability

$$p(X = d | Y = 1) = 8/34$$

Calculate $p(Y = 1)$:

$$p(Y = 1) = \sum_X p(X, Y = 1) = 34/60$$

$$p(X = d, Y = 1) = p(X = d | Y = 1)p(Y = 1)$$

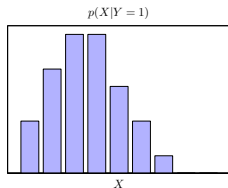
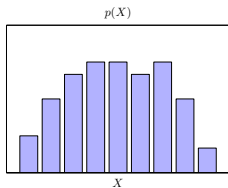
$$p(X, Y) = p(X | Y)p(Y)$$



| Y vs. X | a | b | c | d | e | f | g | h | i | sum |
|---------|---|---|---|---|---|---|---|---|---|-----|
| 2 | 0 | 0 | 0 | 1 | 4 | 5 | 8 | 6 | 2 | 26 |
| 1 | 3 | 6 | 8 | 8 | 5 | 3 | 1 | 0 | 0 | 34 |
| sum | 3 | 6 | 8 | 9 | 9 | 8 | 9 | 6 | 2 | 60 |

$$p(X) = \sum_Y p(X, Y)$$

$$p(X, Y) = p(X | Y) p(Y)$$





- Sum Rule

$$p(X) = \sum_Y p(X, Y)$$

- Product Rule

$$p(X, Y) = p(X | Y) p(Y)$$



Use product rule

$$p(X, Y) = p(X | Y)p(Y) = p(Y | X)p(X)$$

Bayes Theorem

$$p(Y | X) = \frac{p(X | Y)p(Y)}{p(X)}$$

only defined for $p(X) > 0$

and

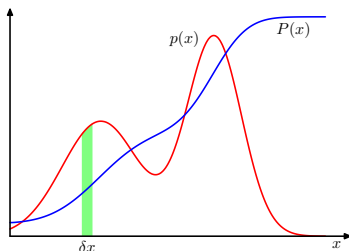
$$p(X) = \sum_Y p(X, Y) \quad \text{(sum rule)}$$

$$= \sum_Y p(X | Y)p(Y) \quad \text{(product rule)}$$



- Real valued variable $x \in \mathbb{R}$
- Probability of x to fall in the interval $(x, x + \delta x)$ is given by $p(x)\delta x$ for infinitesimal small δx .

$$p(x \in (a, b)) = \int_a^b p(x) dx.$$

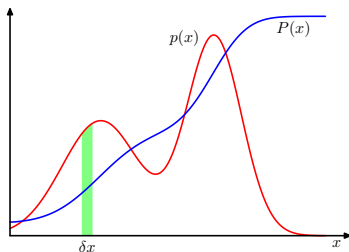


Constraints on $p(x)$

- Nonnegative
- Normalisation

$$p(x) \geq 0$$

$$\int_{-\infty}^{\infty} p(x) dx = 1.$$



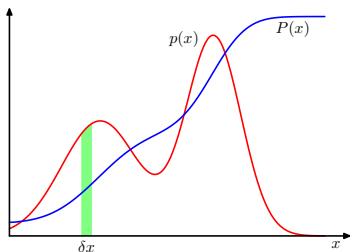
Cumulative distribution function $P(x)$



$$P(x) = \int_{-\infty}^x p(z) dz$$

or

$$\frac{d}{dx}P(x) = p(x)$$



Multivariate Probability Density



- Vector $\mathbf{x} \equiv (x_1, \dots, x_D)^T = \begin{bmatrix} x_1 \\ \vdots \\ x_D \end{bmatrix}$

- Nonnegative

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

- Normalisation

$$\int_{-\infty}^{\infty} p(\mathbf{x}) \, d\mathbf{x} = 1.$$

- This means

$$\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} p(\mathbf{x}) \, dx_1 \dots dx_D = 1.$$

Sum and Product Rule for Probability Densities



- Sum Rule

$$p(x) = \int_{-\infty}^{\infty} p(x, y) \, dy$$

- Product Rule

$$p(x, y) = p(y \mid x) p(x)$$



- Weighted average of a function $f(x)$ under the probability distribution $p(x)$

$$\mathbb{E}[f] = \sum_x p(x)f(x) \quad \text{discrete distribution } p(x)$$

$$\mathbb{E}[f] = \int p(x)f(x) dx \quad \text{probability density } p(x)$$

How to approximate $\mathbb{E}[f]$



- Given a finite number N of points x_n drawn from the probability distribution $p(x)$.
- Approximate the expectation by a finite sum:

$$\mathbb{E}[f] \simeq \frac{1}{N} \sum_{n=1}^N f(x_n)$$

- How to draw points from a probability distribution $p(x)$?
Lecture coming about “Sampling”

Expectation of a function of several variables



- arbitrary function $f(x, y)$

$$\mathbb{E}_x [f(x, y)] = \sum_x p(x) f(x, y) \quad \text{discrete distribution } p(x)$$

$$\mathbb{E}_x [f(x, y)] = \int p(x) f(x, y) dx \quad \text{probability density } p(x)$$

- Note that $\mathbb{E}_x [f(x, y)]$ is a function of y .



- arbitrary function $f(x)$

$$\mathbb{E}_x [f | y] = \sum_x p(x | y) f(x) \quad \text{discrete distribution } p(x)$$

$$\mathbb{E}_x [f | y] = \int p(x | y) f(x) dx \quad \text{probability density } p(x)$$

- Note that $\mathbb{E}_x [f | y]$ is a function of y .
- Other notation used in the literature : $\mathbb{E}_{x|y} [f]$.
- What is $\mathbb{E} [\mathbb{E} [f(x) | y]]$? Can we simplify it?
- This must mean $\mathbb{E}_y [\mathbb{E}_x [f(x) | y]]$. (Why?)

$$\begin{aligned} \mathbb{E}_y [\mathbb{E}_x [f(x) | y]] &= \sum_y p(y) \mathbb{E}_x [f | y] = \sum_y p(y) \sum_x p(x|y) f(x) \\ &= \sum_{x,y} f(x) p(x, y) = \sum_x f(x) p(x) \\ &= \mathbb{E}_x [f(x)] \end{aligned}$$



- arbitrary function $f(x)$

$$\text{var}[f] = \mathbb{E} [(f(x) - \mathbb{E} [f(x)])^2] = \mathbb{E} [f(x)^2] - \mathbb{E} [f(x)]^2$$

- Special case: $f(x) = x$

$$\text{var}[x] = \mathbb{E} [(x - \mathbb{E} [x])^2] = \mathbb{E} [x^2] - \mathbb{E} [x]^2$$



- Two random variables $x \in \mathbb{R}$ and $y \in \mathbb{R}$

$$\begin{aligned}\text{cov}[x, y] &= \mathbb{E}_{x,y} [(x - \mathbb{E}[x])(y - \mathbb{E}[y])] \\ &= \mathbb{E}_{x,y} [xy] - \mathbb{E}[x] \mathbb{E}[y]\end{aligned}$$

- With $\mathbb{E}[x] = a$ and $\mathbb{E}[y] = b$

$$\begin{aligned}\text{cov}[x, y] &= \mathbb{E}_{x,y} [(x - a)(y - b)] \\ &= \mathbb{E}_{x,y} [xy] - \mathbb{E}_{x,y} [xb] - \mathbb{E}_{x,y} [ay] + \mathbb{E}_{x,y} [ab] \\ &= \mathbb{E}_{x,y} [xy] - b \underbrace{\mathbb{E}_{x,y} [x]}_{=\mathbb{E}_x[x]} - a \underbrace{\mathbb{E}_{x,y} [y]}_{=\mathbb{E}_y[y]} + ab \underbrace{\mathbb{E}_{x,y} [1]}_{=1} \\ &= \mathbb{E}_{x,y} [xy] - ab - ab + ab = \mathbb{E}_{x,y} [xy] - ab \\ &= \mathbb{E}_{x,y} [xy] - \mathbb{E}[x] \mathbb{E}[y]\end{aligned}$$

- Expresses how strongly x and y vary together. If x and y are independent, their covariance vanishes.

Covariance for Vector Valued Variables



- Two random variables $\mathbf{x} \in \mathbb{R}^D$ and $\mathbf{y} \in \mathbb{R}^D$

$$\begin{aligned}\text{cov}[\mathbf{x}, \mathbf{y}] &= \mathbb{E}_{\mathbf{x}, \mathbf{y}} [(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{y}^T - \mathbb{E}[\mathbf{y}^T])] \\ &= \mathbb{E}_{\mathbf{x}, \mathbf{y}} [\mathbf{x} \mathbf{y}^T] - \mathbb{E}[\mathbf{x}] \mathbb{E}[\mathbf{y}^T]\end{aligned}$$